Pautas metodológicas para investigaciones cualitativas y cuantitativas en ciencias sociales y humanas

Mario Yapu (Coordinador)

Denise Y. Arnold
Alison L. Spedding P.
Rodney Pereira M.

Universidad de Postgrado para la Investigación Estratégica en Bolivia (U-PIEB)

La Paz, 2013

Yapu, Mario
Pautas metodológicas para investigaciones cualitativas y cuantitativas en ciencias sociales y humanas / Mario Yapu; Denise Y. Arnold; Alison L. Spedding P.; Rodney Pereira M. -- 4 ed. -- La Paz: Fundación PIEB, 2013.
xxix; 309 p.; cuadrs.; tbls.; grafs; 23 cm. -- (Serie Metodológica U-PIEB)

D.L.: 4-1-2141-3
ISBN: 978-99954-57-66-2 : Encuadernado

MÉTODOS DE INVESTIGACIÓN / MÉTODOS DE INVESTIGACIÓN-CIENCIAS SOCIALES / ORGANIZACIÓN DE LA INVESTIGACIÓN / MÉTODOS DE INVESTIGACIÓN-CUALITATIVA / MÉTODOS DE INVESTIGACIÓN-CUANTITATIVA / RECOLECCIÓN DE DATOS / INVESTIGACIÓN DE CAMPO / INVESTIGACIÓN EMPÍRICA / ANÁLISIS DE DATOS / CONTEXTO DE LA INVESTIGACIÓN / ESTRATEGIA METODOLÓGICA / ESTADO DE LA CUESTIÓN

1. título 2. serie

D.R.© Fundación PIEB, mayo 2006
Cuarta edición, octubre 2013
U-PIEB. Universidad de Postgrado para la Investigación Estratégica en Bolivia
Edificio Fortaleza. Piso 6. Oficina 601
Avenida Arce 2799, esquina calle Cordero
Teléfonos: 2432582 - 2431866
Fax: 2435235
Correo electrónico: upieb@upieb.edu.bo
Página web: www.upieb.edu.bo
Casilla 12668
La Paz, Bolivia

Edición: Ana Rebeca Prada

Diseño gráfico de cubierta: PIEB

Diagramación: Alfredo Revollo Jaén

Impresión:

Impreso en Bolivia
Printed in Bolivia

CAPÍTULO TRES

Metodologías cuantitativas, operacionalización de la investigación, recolección y análisis de datos

Rodney Pereira M.

Introducción

Desde sus inicios, el PIEB ha estimulado tanto la investigación como la formación de investigadores, para lo cual ha realizado talleres de actualización y ha publicado el libro *Guía para la formulación y ejecución de proyectos de investigación* (Barragán 2007) que al presente se encuentra en su cuarta edición y que se ha convertido en una referencia importante para el investigador social y para docentes que guían a estudiantes en sus proyectos de investigación.

La *Guía* muestra un panorama general de los diferentes aspectos y momentos para elaborar un proyecto de investigación, incorporando temas que cubren desde la formulación de los problemas, las técnicas de recolección de la información, las opciones para probar la validez interna y externa, los estilos de redacción de informes finales, hasta la forma de presentación de las investigaciones.

Sin embargo, la amplitud de los temas tratados en la *Guía* no ha permitido profundizar algunos aspectos de los componentes metodológicos de la investigación, en particular, los referidos a la estrategia metodológica con énfasis cuantitativo.

En tal sentido y con el propósito de constituirse en un documento complementario a la *Guía*, se ha elaborado el presente texto que tiene el objetivo de presentar en forma didáctica y comprensible los elementos más importantes de la investigación cuantitativa con aplicaciones al área social, considerando que el lector no tiene una formación estadística. Debe tenerse en cuenta que, sobre este tema, las técnicas cuantitativas han mostrado una significativa evolución en los últimos años, desplegando niveles de sofisticación que requieren para su aplicación una sólida base matemática, de estadística básica y teoría de probabilidades.

Por lo tanto, hacer que un texto, por un lado, sea accesible a investigadores sociales que no necesariamente tienen formación cuantitativa y, por otro, aborde de manera adecuada los principales aspectos que hacen a la estrategia metodológica desde la perspectiva cuantitativa, implica fijar un alcance que se definió en función del perfil del investigador social que en el caso boliviano se orienta a buscar una mejor compresión de los fenómenos sociales en espacios regionales determinados.

Bajo la anterior consideración, el texto aborda en forma ordenada las diferentes fases del diseño metodológico para una investigación cuantitativa: en el primer apartado del capítulo se efectúan algunos apuntes sobre la incorporación de los aspectos cuantitativos en las investigaciones sociales en Bolivia y, a manera de introducción, se presentan las características esenciales de las investigaciones de tipo cuantitativo y cualitativo y sus principales diferencias, así como las fases del proceso de la investigación cuantitativa.

El segundo apartado, parte central del texto, se refiere a la estrategia metodológica en la cual se aborda los diferentes componentes de la misma. Se parte de algunas definiciones básicas, luego se realiza una revisión de los tipos de investigación desde la perspectiva cuantitativa y el significado de la validez interna, externa y de la evaluación, posteriormente se exponen los aspectos de la operacionalización que comprende la forma de cómo los conceptos se transforman en distintos tipos de variables cuantificables, su forma de medición y la operativización de las hipótesis. Se continúa con la exposición de los procedimientos de selección de las unidades de observación, para lo cual se realiza una introducción al muestreo destacando aspectos como el marco muestral, los tipos de muestreo, el tamaño de la muestra y los errores muestrales.

En un tercer apartado se incorporan los aspectos referidos a la recolección de datos, desde el diseño del cuestionario, hasta las formas de entrevista, el trabajo de campo y la codificación.

Finalmente, se aborda la fase del análisis de datos, en la que se incorporan temas como la elaboración de tablas de frecuencias y de contingencia, las formas de presentación gráfica y las medidas para el análisis descriptivo de tendencia central, dispersión y de forma.

Cada apartado y subtítulo se acompaña de ejemplos que permiten clarificar las definiciones y fórmulas. En muchos casos, para la construcción de los ejemplos se utilizó el paquete SPSS por su versatilidad, fácil manejo y debido a que las bases de datos utilizadas se encuentran en este formato.

Por los alcances mencionados y sus características, se puede calificar al presente capítulo como introducción a las estrategias cuantitativas de investigación social. Aspectos tales como los amplios temas referidos a la inferencia o estimación estadística y pruebas de hipótesis no están incorporados debido a que su abordaje requiere conocimientos sobre la teoría de probabilidades, y porque su desarrollo requeriría un amplio espacio en el texto. Sin embargo, la bibliografía sobre este tema es abundante y en algunos casos de fácil acceso.

En muchos casos, los ejemplos fueron construidos en base a la información proveniente de las Encuestas de Hogares y Calidad de Vida, proporcionada gentilmente por funcionarios del INE, los que merecen una mención especial. Asimismo, debo destacar la labor de revisión y sugerencias del Dr. Mario Yapu, que ayudaron a mejorar la elaboración del presente texto.

1. Apuntes de investigación social y metodologías cuantitativas

1.1. Contexto de investigaciones

La forma de abordaje a los temas de investigación social en Bolivia ha estado signada en gran medida por la fuerte influencia de los denominados trabajos de tesis que se realizan en las universidades como requisito para la titulación, los cuales en general se ciñen a formatos rígidos sin distinción de las áreas de especialización. Esta exigencia influye para que de manera mecánica se incorporen en las tesis componentes que en muchos casos no son pertinentes para la temática elegida; esto sucede en particular en las disciplinas del área social.

Salvo excepciones, la mayor parte de estos trabajos contienen objetivos muy generales, hipótesis obvias y numerosas, marcos teóricos sobreabundantes y débiles propuestas metodológicas, aspectos que afectan la operacionalización de los conceptos, la definición de las variables, la recolección de información y su análisis, lo cual determina la falta de coherencia entre objetivos, hipótesis, marcos teóricos, metodología y resultados.

Además, también influyen en el ámbito universitario los sesgos y "modas". Por ejemplo, en las áreas sociales se han creado mitos sobre las técnicas cuantitativas, a veces con argumentos de una imagen de complejidad, a veces por preconceptos sobre la validez de las mismas. En algunos casos, por su utilización en general como una moda, se las emplea sin una comprensión adecuada, como un requisito más.

Por otro lado, en disciplinas sociales como la economía aparece un fuerte sesgo hacia el sobreuso de técnicas cuantitativas asociadas a los métodos econométricos, cuyas tareas se ven facilitadas con el desarrollo de paquetes computacionales; sin embargo, el análisis de resultados continúa siendo débil y en muchos casos estas "cajas negras" definen la validez de la investigación. Adicionalmente, se margina aquello que no se puede medir, lo cual influye en la no incorporación de métodos cualitativos que pueden complementar este tipo de investigaciones.

En el sector público, las investigaciones son escasas; abundan los diagnósticos en muchos casos recargados de información estadística que no permite relevar los aspectos más importantes. Si bien se utilizan cuadros y gráficos de series temporales, éstos en su mayor parte no se acompañan con indicadores de tendencia central y de dispersión o análisis de correlación y regresión u otros métodos estadísticos.

Existen excepciones como la Unidad de Análisis de Políticas Sociales y Económicas (UDAPE), en cuyas investigaciones, diagnósticos y proyecciones el uso de métodos cuantitativos es un aspecto central. Asimismo, el Instituto Nacional de Estadística (INE) incorpora cada vez más en sus distintas publicaciones métodos cuantitativos para la presentación de indicadores y análisis de resultados. Se destaca el Fondo de Estudios del Programa de Mejoramiento de las Encuestas y Medición de las Condiciones de Vida (MECOVI) dependiente del INE, que ha auspiciado un conjunto de investigaciones de buena calidad relacionadas con aspectos de la pobreza, las mismas que se publican en la revista *Estadísticas y análisis*.

En los últimos años, gracias a la apertura de espacios como el PIEB, fundaciones, centros de investigación y redes de investigación internacional que promueven la realización de estudios sobre distintos tópicos de la realidad nacional con énfasis en los aspectos sociales, e incentivan —como el caso de la U-PIEB— una mayor calificación de los investigadores a través de cursos formales, talleres y seminarios, se ha incidido en la cualificación de las investigaciones en todos sus aspectos y, en particular, en el tratamiento y análisis de la información con la incorporación de métodos cuantitativos que, si bien aun no son muy rigurosos, enriquecen la investigación social. La revisión de una muestra de investigaciones realizadas en el marco de las convocatorias que realiza el PIEB ratifica esta aseveración.

Por ejemplo, en el enfoque metodológico de la investigación *Tarija en los imaginarios urbanos* (Lea Plaza/Vargas/Paz 2003) se considera la complementariedad de los métodos cuantitativos y cualitativos, "pues permiten explorar y comprender un problema desde varias dimensiones:

la representatividad, profundidad y especificidad". El estudio define tres ámbitos de indagación: evocaciones históricas, representaciones actuales e idealizaciones y proyecciones, cada uno con sus propias categorías. Los instrumentos utilizados fueron entrevistas en profundidad para las evocaciones, las encuestas para las representaciones, y los grupos focales para las idealizaciones.

Para la realización de la encuesta se estableció una muestra de carácter probabilístico que se aplicó a 313 personas; la mitad se efectuó en la denominada zona tradicional de la ciudad de Tarija y la otra en la zona de inmigrantes. Los resultados de la encuesta se muestran utilizando representaciones gráficas y frecuencias relativas expresadas en porcentajes para la explicación de las percepciones sobre el tamaño, población, percepciones sensoriales, los centros de la ciudad, lugares representativos, calidad de vida y otros.

En esta investigación se destaca la incorporación de aspectos cuantitativos mediante la utilización de variables cualitativas. Sin embargo, no se incorpora el método para definir el tamaño de la muestra, el nivel de confianza y margen de error, elementos que hacen a la validez externa. Asimismo, no se explicita el tipo de muestreo utilizado, aleatorio simple, sistemático, estratificado, etc. Los resultados van más allá de las percepciones en cuanto se presentan estados de situación. La explotación de la información primaria podría haberse enriquecido con algunas tablas de contingencia que permitan un primer nivel de cruce de variables.

La investigación *La permanente construcción de lo cruceño* (Peña/Barahona/Rivero/Gaya 2003) busca responder a las siguientes preguntas: ¿Qué es lo cruceño?, ¿qué es la cruceñidad?, y ¿qué es el cruceñismo? Si bien no presenta explícitamente el diseño metodológico, éste se encuentra en la introducción de manera sucinta. La investigación tiene un carácter de revisión histórica y bibliográfica para indagar acerca de la identidad cruceña, la misma que se complementa con una encuesta aplicada a una muestra de la población.

En los aspectos cuantitativos, se utilizó una encuesta de 26 preguntas de carácter cualitativo. En base a información del INE del Censo de 1992 y con un nivel de confianza de 95%, se definió un tamaño de muestra de 935 personas. De esta muestra, 50% estuvo constituido por grupos de población preestablecidos: políticos, diputados concejales, empresarios, universitarios, entre otros. El restante 50% fue población en general. Para el análisis de datos se utilizaron gráficas y tablas de frecuencias absolutas y relativas y de contingencia ordenadas por grupos de edad y lugar de origen.

En este caso, se observa un mejor aprovechamiento de la información. En cuanto a la muestra, no se especifica la metodología empleada para definir el tamaño y tampoco se aclara el tipo de muestreo realizado. Una primera impresión es que éste es aleatorio, sin embargo, al aplicarse a grupos preestablecidos sin definir la forma de selección, el carácter aleatorio quedaría cuestionado y, por lo tanto, la validez externa.

La investigación sobre *Calidad de vida en Potosí* (Tapia/Quintana/Ance/Morales 2002) tiene el propósito de incorporar los aspectos ambientales en la medición de la calidad de vida y su incidencia en el bienestar de la población de la ciudad de Potosí. Incorpora un buen marco metodológico en el que se establecen claramente las unidades de análisis, las variables y los instrumentos, el tipo de muestreo y el tamaño de la muestra con sus respectivos intervalos de confianza y márgenes de error.

En este estudio se destaca la metodología para la medición de la calidad de vida a través de la aplicación de los indicadores de necesidades básicas insatisfechas (NBI) a las familias seleccionadas en la muestra. Para el ordenamiento y análisis de la información se utilizó el paquete estadístico SPSS, permitiendo una mayor fiabilidad en el cálculo de los diferentes indicadores utilizados.

1.2. Investigación cuantitativa y cualitativa

1.2.1. Diferencias principales

En el marco de las ciencias sociales se pueden distinguir dos perspectivas metodológicas para la investigación: la cuantitativa y la cualitativa. Ambas difieren tanto en la estrategia de recolección de la información como en su análisis y en su base epistemológica.

La investigación cuantitativa epistemológicamente está vinculada al empirismo y al positivismo lógico, que tienden a estudiar los eventos desde el exterior, mientras que la investigación cualitativa se apoya en la fenomenología, el interaccionismo simbólico y el historicismo, y se interesa en la perspectiva de los actores y su experiencia vivida, (Salman 2003: 94).

El objetivo de la investigación cuantitativa es la medición de los hechos, la demostración de relaciones de causalidad entre variables o prueba de las hipótesis, y la generalización de los resultados mediante un proceso de inducción (inferencia). Por su parte, la investigación cualitativa se orienta a la descripción, interpretación y comprensión de la conducta humana o de un determinado grupo en el mismo ambiente donde se desenvuelve.

La recolección de la información desde la perspectiva cuantitativa se realiza de manera estructurada y sistemática, a través de encuestas que permiten la medición de variables en una relación diferenciada entre el observador y el observado, a fin de lograr la objetividad de la información. En la investigación cualitativa la información se obtiene mediante un proceso interactivo y participativo en base a la definición de dominios de interés, planes para observar y explorar sin considerar variables operacionales.

El análisis cuantitativo se preocupa por magnitudes, proporciones, datos agregables, en definitiva, por la cuantificación de los aspectos de la realidad social, de sus relaciones de causalidad con el propósito de encontrar regularidades o leyes sociales, mientras que la perspectiva cualitativa analiza la conducta humana desde el punto de vista del actor, mediante la interpretación de su lenguaje, acciones e interrelaciones sociales con objeto de comprender el mundo social.

En el siguiente cuadro se resumen las principales características de las dos perspectivas metodológicas:

	Método cuantitativo	Método cualitativo
Base epistemológica	Empirismo, positivismo lógico, funcionalismo	Fenomenología, interaccionismo simbólico, historicismo
Énfasis	Medición objetiva de los hechos sociales; demostración de la causalidad y generalización de los resultados	Descripción y comprensión interpretativa de la conducta humana en el ambiente donde el individuo o grupo social actúa
Recolección de información	Estructurada y sistemática	Flexible, es un proceso interactivo, marcado por el desarrollo de la investigación
Análisis	Cuantificar la realidad social, las relaciones de causalidad y su intensidad	Interpretar el lenguaje, los discursos, acciones y estructuras latentes
Alcance	Búsqueda de regularidades o leyes de comportamiento social	Búsqueda cualitativa de significados de la acción humana

Fuente: Cea D'Ancona 1999: 46.

Pese a las diferencias y el debate sobre estos tipos de investigación, hoy en día la mayoría de los investigadores sociales reconoce que en la práctica la mayor parte de las investigaciones contienen elementos cuantitativos y cualitativos y que ambos son complementarios (Barragán 2003: 9).

Son dos las razones para esta complementariedad: la primera, que la investigación cualitativa ha demostrado tener métodos y procedimientos válidos importantes y nada débiles; la segunda, que la investigación

cuantitativa no es tan objetiva porque utiliza una serie de elementos cualitativos. Cada vez se reconoce más que dimensiones de la realidad social no pueden ser cuantificadas y que lo cualitativo es necesario para el análisis cuantitativo. Es el caso de las correlaciones entre variables numéricas que, para su mejor interpretación, requieren aspectos cualitativos (Salman 2003: 102).

Por lo tanto, la dicotomía entre lo cuantitativo y cualitativo no es tal a la hora de emprender una investigación social que en general tiene características multidimensionales (Salman 2003: 104).

1.2.2. El método cuantitativo

Se considera al método cuantitativo como representante del método científico; éste consiste en un proceso secuencial, continuo y circular mediante el cual se conectan la deducción, operacionalización, interpretación e inducción a través de cuatro componentes: teorías, hipótesis, observaciones y generalizaciones empíricas (Wallace 1976).

De acuerdo a este esquema, el proceso de investigación cuantitativa se realiza de la siguiente manera:

a) Se parte de una teoría general o particular cuyas proposiciones están lógicamente interconectadas y de las que pueden deducirse uniformidades empíricas (Merton 1983: 56).

b) De esta teoría se extrae, para su contrastación empírica, una o varias hipótesis concretas, las cuales representan respuestas probables o sospechas a las interrogantes que el investigador se plantea.

c) Las hipótesis se formulan como proposiciones en las que se afirma la existencia o no de una relación esperada entre al menos dos variables.

d) Las proposiciones y conceptos que conformas las hipótesis se concretan en variables e indicadores que posibiliten su contrastación empírica; esta fase requiere la definición de las unidades de observación.

e) Seleccionadas las variables, indicadores y el ámbito de la investigación, se procede a la recolección de la información en base a una estrategia metodológica elegida por el investigador.

f) La información se procesa y analiza con métodos estadísticos y se procede a su interpretación para su generalización empírica.

g) Las generalizaciones empíricas se contrastan con las hipótesis de investigación mediante un proceso inferencial conocido como contrastación de hipótesis.

h) Si los datos corroboran la hipótesis, se comprueba la validez externa de la investigación; si la investigación pretende comprobar una teoría, ésta se valida. En caso contrario, se rechazan las hipótesis o la teoría por los hallazgos empíricos y el proceso de investigación comienza de nuevo. Debe tenerse presente que "cualquiera que sea el método para la verificación de las hipótesis, los resultados nunca son ciertos sino aproximaciones en términos de probabilidad" (Goode y Hatt 1975: 87).

En resumen, el método cuantitativo involucra un proceso de conocimiento deductivo e inductivo. La fase deductiva comprende el marco teórico, la formulación de hipótesis, su operacionalización y la recolección de datos. Por su parte, la fase inductiva se inicia a partir del análisis e interpretación de los datos, su generalización empírica, la verificación de las hipótesis y el testeo de una teoría.

Bajo este marco, en los siguientes apartados se abordará la estrategia metodológica desde la perspectiva cuantitativa para la formulación y ejecución de proyectos de investigación social.

2. La estrategia metodológica

En el ámbito de la investigación social es necesario precisar el alcance del proyecto y del diseño o estrategia de investigación. El proyecto involucra todos los aspectos relativos al proceso de investigación hasta el detalle económico, organizativo y temporal de sus diferentes fases e incorpora como uno de sus componentes la estrategia para alcanzar los objetivos.

Alvira define a la estrategia como un "plan global de la investigación que integra, de modo coherente y adecuadamente correcto, técnicas de recogida de datos a utilizar, análisis previstos y objetivos" (1989: 85). La *Guía para la formulación y ejecución de proyectos de investigación* del PIEB considera a la estrategia como el conjunto de operaciones y decisiones metodológicas para la obtención y posterior análisis de la información que debe orientarse por los objetivos, hipótesis y por las características del tipo de investigación que se pretende realizar, ya sea cuantitativo o cualitativo (Barragán 2003: 87). También destaca que la estrategia metodológica constituye el cincuenta por ciento de un proyecto de investigación y que los proyectos pecan en algunos casos por no otorgar a este aspecto el lugar que tiene y debe tener.

Por estas razones, la estrategia metodológica constituye un aspecto central en los proyectos de investigación, tanto en su fase de formulación como de ejecución. Una buena idea de investigación puede no llegar a resultados significativos cuando el diseño metodológico no es el adecuado y, en algunos casos, llevar a resultados erróneos.

Componentes básicos de un proyecto de investigación

Formulación del problema de investigación
- Definir objetivos
- Elaborar el marco teórico: revisión bibliográfica e indagaciones exploratorias

Diseño de la investigación
- Experimental, no experimental

Operacionalización del problema
- Formular hipótesis
- Detectar las variables
- Definir conceptualmente las variables
- Definir operacionalmente las variables
- Delimitar las unidades de análisis

Selección de muestra
- Determinar el universo
- Tamaño y error muestral
- Procedimiento de selección

Recolección de datos
- Diseño de formularios
- Instrumento de medición
- Codificación de datos
- Procesamiento de datos

Análisis de datos
- Estadístico, gráfico y otros

Como se observa en el cuadro, la estrategia presupone que se han definido el tema, el marco teórico, los objetivos, las hipótesis y que se ha efectuado la revisión bibliográfica. Estos aspectos son tratados con amplitud en la mencionada *Guía* del PIEB (Barragán 2003). En lo que sigue y en el marco de los objetivos del presente texto, se abordarán cada uno de los componentes de la estrategia metodológica para aquellas investigaciones que destaquen aspectos cuantitativos.

2.1. Tipos de investigación

No existe una forma estándar para realizar una investigación; es como en la construcción, en la que existen varios planos y éstos están divididos en modelos especiales (Namakforoosh 2000: 85). Entre los

tipos de investigación más generalizadas, se pueden mencionar las no experimentales y las experimentales. Entre las primeras se destacan las exploratorias, descriptivas, causales y temporales.

2.1.1. Investigación exploratoria

Su objetivo principal es captar una perspectiva general del problema. Este tipo de estudios ayudan a dividir un problema muy grande y llegar a subproblemas más precisos que ayuden a definir las hipótesis. También es útil para incrementar el grado de conocimiento del investigador respecto al objeto de estudio.

2.1.2. Investigación descriptiva

Este tipo de investigación sirve para saber quién, dónde, cuándo, cómo y por qué. Se usa cuando el objetivo de la investigación es describir las características de ciertos grupos. Por ejemplo, elaborar el perfil de los pobres, calcular la proporción de personas de una población específica con características particulares.

2.1.3. Investigación causal

En esencia, todas las investigaciones científicas tienen una idea de causalidad, es decir, se espera que una variable independiente produzca ciertos cambios en la variable dependiente. Sin embargo, saber que existe una relación entre dos variables no es suficiente para la investigación científica.

Hay cierta diferencia entre el pensamiento científico y el sentido común. Cuando se dice que X afecta Y, implica que si cambia X cambiará Y, mientras que desde el punto de vista científico X puede ser uno de los varios casos en que puede influir sobre la variable Y. Otra diferencia: el sentido común cree que la relación es determinística, mientras que el conocimiento científico dice que la relación es probabilística. Finalmente, el pensamiento científico de causalidad establece que nunca se puede decir con certeza que X puede ser de Y, sino que siempre se hace la inferencia de que la relación existe (Namakforoosh 2000: 93).

2.1.4. Investigaciones temporales o longitudinales

Recolectan datos a través del tiempo en puntos o periodos específicos, para analizar cambios, determinantes y consecuencias. Por ejemplo, analizar cómo ha evolucionado la cobertura escolar en un departamento durante cinco años. Estas investigaciones pueden ser de tres tipos: de tendencia, de análisis de cohortes y de panel:

a) *De tendencia*. Analizan cambios a través del tiempo dentro de una población. Por ejemplo, interesa conocer los cambios en los movimientos migratorios durante diez años.

b) *De evolución de cohortes*. Examinan cambios a través del tiempo en grupos específicos. Por ejemplo, interesa conocer el número y las causas de deserción de estudiantes que en 1992 ingresaron a la escuela y que en 2004 concluyeron el bachillerato.

c) *De panel*. Son similares a los diseños de cohorte sólo que en el grupo los mismos elementos son medidos en los diferentes momentos en el tiempo.

2.1.5. Investigación experimental

En este tipo de investigación se tiene control directo en al menos una variable y se puede manipular por lo menos una variable independiente. Los componentes de este tipo de investigación consisten en: comparación, manipulación y control. La *comparación* es una operación requerida para demostrar que dos variables están correlacionadas. *Manipulación* significa que un cambio inducido en X será seguido por un cambio en Y. El *control* se refiere a la interrogante de si la variable independiente es una causa para modificaciones en la variable dependiente.

En las ciencias sociales los experimentos están limitados en la medida que es difícil reproducir condiciones ideales que se podrían realizar en un laboratorio, por lo que los análisis de causalidad deberán realizarse con cruzamiento de datos, correlaciones y otras técnicas estadísticas. En la actualidad existen varias opciones para simular experimentos[1].

2.2. Evaluación y validez de las formas de investigación

Existen varios criterios a seguir a la hora de evaluar un diseño de investigación, sin embargo, el criterio fundamental es que el diseño se adecue a los objetivos. Si esto no se logra, no se justifica realizar las actividades planificadas. Si se cumple este requisito, se pueden considerar cuatro criterios de validez para la investigación cuantitativa: la interna, la externa, la de constructo y la estadística.

---

1 Para una ampliación de este tema, se recomienda consultar *Metodología de la investigación* de Hernández, Fernández y Baptista (1999), Cap. 6.

2.2.1. Validez interna

Desde la perspectiva cuantitativa, existe validez interna cuando las relaciones de causalidad establecidas entre las variables dependientes e independientes no permiten explicaciones alternativas o éstas pueden ser controladas. La comprobación de este tipo de validez es prioritaria sobre todo en diseños explicativos. Si este criterio de validez no se satisface, los resultados de la investigación serían cuestionables y siempre existirán posibles explicaciones alternativas a las relaciones observadas (Cea D'Ancona 1999: 117).

Por lo tanto, el mayor o menor grado de validez interna depende del control de explicaciones alternativas a las relaciones observadas, es decir, del número de variables perturbadoras cuya influencia se ha controlado o neutralizado (ver Ejemplo No. 3).

2.2.2. Validez externa

Representa la posibilidad de generalización de los resultados de una investigación a la población. En la práctica, dadas las limitaciones (por presupuesto y tiempo) de estudiar el todo, se analiza una muestra que deberá ser representativa de una población y los resultados deberán ser factibles de generalización hacia esta población. La representatividad de la muestra está subordinada al procedimiento de selección de los elementos de la muestra.

2.2.3. Validez de constructo

Se utiliza para considerar el grado de adecuación conseguido en la medición de los conceptos centrales de la investigación. Cualquier concepto permite distintas posibilidades de medición, a su vez, es difícil que cualquier operacionalización de un concepto cubra todas las dimensiones del mismo. Por lo tanto, es necesario una operacionalización precisa de los conceptos; ello contribuye a reducir la duda si se hubiera alcanzado los mismos resultados con una operacionalización distinta (Cea D'Ancona 1999: 120).

Ejemplo No. 1

Se pueden considerar dos definiciones para medir el desempleo:
Las personas mayores de 15 años que no trabajan
Las personas mayores de 15 años que buscan trabajo y no lo encuentran

Si se mide el desempleo por la primera definición, se incluirá personas que no desean o no pueden trabajar (estudiantes, jubilados, amas de casa). Por la segunda definición, la medición sólo incluirá a las personas que desean trabajar. En ambos casos los resultados serán distintos debido a que se consideran dos conceptualizaciones sobre el desempleo.

2.2.4. Validez estadística

Está relacionada con el poder, adecuación y fiabilidad de la técnica de análisis de datos aplicada. Esta validez utiliza técnicas analíticas de la significación estadística de los resultados y la posibilidad de generalización de los hallazgos obtenidos de la muestra. Antes de utilizar una técnica analítica concreta, ha de comprobarse si la información recabada satisface los supuestos exigidos para la práctica de la técnica de análisis elegida.

2.2.5. La fiabilidad

Se refiere a la capacidad de obtener resultados consistentes en mediciones sucesivas de un mismo fenómeno, es decir, los resultados logrados en mediciones repetidas de un mismo concepto deben ser iguales para que la medición se estime fiable. Una forma de comprobar la fiabilidad consiste en aplicar el mismo procedimiento de medición en diferentes momentos para observar si se obtienen resultados similares en las distintas mediciones del concepto.

2.3. El proceso de operacionalización

El término *operacionalización* se emplea para denotar las formas de medición de los conceptos. En el proceso de operacionalización se deben diferenciar dos nociones fundamentales: la conceptualización y la medición. La conceptualización hace referencia al proceso mediante el cual se clasifican las ideas o las proposiciones teóricas, y la medición es el proceso que vincula las operaciones físicas de medición con las operaciones matemáticas de asignar números a objetos (Cea D'Ancona 1999: 123). Por lo tanto, la operacionalización implica un triple nexo que relaciona los conceptos teóricos con las operaciones físicas de medición y éstas con los símbolos matemáticos.

La medición se define como la asignación de números a objetos o a hechos de acuerdo con determinadas reglas (Stevens 1951). Esta definición se ha revisado para el caso de las ciencias sociales, en las que se considera a la medición como el proceso de vincular conceptos abstractos a indicadores empíricos (Carmines y Zeller 1979: 10).

La generalidad de los conceptos constituyen variables latentes, hipotéticas y no directamente observables; para su concreción se requiere la traducción del concepto teórico a indicadores, a variables empíricas que midan las propiedades enmarcadas en el concepto (Cea D'Ancona 1999: 125).

De acuerdo a Lazarsfeld, se distinguen las siguientes fases en el desarrollo de la operacionalización:

a) Representación teórica del concepto.

b) Especificación del concepto descomponiéndolo en las distintas dimensiones o aspectos relevantes que engloba.

c) Selección de una serie de indicadores (variables empíricas) para las dimensiones elegidas.

d) Asignación de un peso o valor a cada indicador de acuerdo a su importancia.

e) A partir de estos valores, se confecciona un índice, una medida común que agrupe a varios indicadores de una misma dimensión conceptual operacionalizada numéricamente (1973).

En el proceso de operacionalización se debe tener en cuenta los siguientes aspectos:

a) Entre los indicadores y conceptos debe haber plena correspondencia.

b) Los indicadores pueden materializarse en formas diversas (preguntas en un cuestionario, entrevista abierta, registro de una conducta observada), dependiendo de la técnica de obtención de información que se hubiese seleccionado.

c) En la operacionalización se asumen márgenes de incertidumbre. La relación entre los indicadores y el concepto que se trata de medir siempre será supuesta, nunca plenamente cierta.

Dado que la operacionalización implica la asignación de formas de medida a los conceptos, es necesario definir previamente los elementos para efectivizar este proceso.

2.3.1. Las variables

Una variable es cualquier cualidad o característica de un objeto o evento que contenga atributos que permitan su clasificación. Los atributos son los distintos valores o categorías que componen la variable.

Una primera clasificación de las variables es aquella que distingue entre variables cuantitativas y cualitativas. Las cuantitativas toman directamente la forma de números y pueden ser discretas o continuas. Las variables discretas surgen de un proceso de conteo y se conforman por números enteros, por ejemplo, el número de hijos de la familia, el número de cuartos de una vivienda.

Las variables continuas se originan mediante un proceso de medición y adoptan valores numéricos enteros, decimales o fraccionales de la escala real de números. Por ejemplo, la estatura puede tomar cualquier valor dentro de un intervalo; dependiendo de la precisión del instrumento de medición puede establecerse la estatura en metros, centímetros, milímetros y micras (Berenson/Levine 1992: 9).

Las variables cualitativas o categóricas adoptan atributos que no se expresan directamente en forma numérica, por ejemplo, sexo (hombre, mujer), estado civil (soltero, casado, viudo, divorciado, separado), tipo de ocupación (asalariado, independiente).

Variables según su función en la investigación:

Atendiendo a este criterio, las variables cuantitativas y cualitativas se clasifican en independientes, dependientes, perturbadoras y de control.

Las variables independientes o explicativas son aquellas que influyen en las variables dependientes. En general, son parte de las hipótesis e indican posibles causas para la modificación de la variable en la que se centra el interés de la investigación.

Las variables dependientes son aquéllas cuyos atributos numéricos o categóricos (cualitativos) dependen de los valores que tomen las variables independientes.

Ejemplo No. 2

Una investigación tiene el propósito de determinar las variables que inciden en la fecundidad de la mujer. En este caso, la fecundidad será la variable dependiente y las posibles causas de la fecundidad serán las variables independientes, como el nivel de educación de la madre, la edad, su estado civil, su salud, entre otras.

Con el objeto de mejorar los niveles de ingreso de las personas, se desea conocer las variables que inciden sobre el ingreso, en tal caso, el ingreso personal será la variable dependiente y sus posibles determinantes, como el nivel de educación, capacitación, experiencia, edad, tipo de ocupación, serán las variables independientes.

Las variables perturbadoras y de control son aquellas que contribuyen a explicaciones alternativas a la propuesta en la investigación. Esta situación se debe a que entre una variable independiente y una variable dependiente cabe la posibilidad de que existan otras variables que median en la relación. Si el efecto de esas variables se controla, antes o después de haber recogido los datos, dejan de ser perturbadoras y se convierten en variables de control. En toda investigación se debe procurar controlar el mayor número posible de variables perturbadoras: ello favorece a la validez interna de la investigación.

Ejemplo No. 3

Una de las variables independientes que explica la fecundidad puede ser el nivel de educación, sin embargo, esta variable puede estar mediada por aspectos tales como el acceso a la educación (la existencia de escuela, la distancia), el nivel de pobreza, la condición laboral de la madre (si trabaja o no y el tipo de ocupación que desempeña). De no considerarse estas variables, puede surgir una explicación alternativa sobre los determinantes de la fecundidad. Cuando se incorporan estos aspectos a la investigación se habrá definido un conjunto de variables de control que afectan diferencialmente en el nivel de educación de la madre.

Se considera que el rendimiento académico (variable dependiente), medido en una escala de calificación, depende de las horas semanales destinadas al estudio (variable independiente). Para que esta relación tenga validez interna se debe controlar el efecto de otras variables que median en la relación observada, como el coeficiente de inteligencia y la asistencia a clases.

(Continúa en la página siguiente)

(Viene de la página anterior)

Luego, para concluir que existe una relación entre el rendimiento académico y las horas de estudio, previamente se tendrá que haber comprobado que independientemente del coeficiente de inteligencia y la asistencia a clases se presenta la relación. Esto se logra seleccionando un conjunto de estudiantes que en promedio tiene la misma calificación y se selecciona un grupo que tenga en promedio los mismos coeficientes de inteligencia y asistencia a clases y otro donde estas variables muestren distintos valores; se comparan ambos grupos y si se comprueba que no existe relación entre estas variables se concluye que la relación entre rendimiento académico y horas de estudio es válida, de lo contrario, la relación no se sostiene y cabe la posibilidad de una explicación alternativa sobre el rendimiento académico.

Variables según el nivel de abstracción:

De acuerdo a este criterio, las variables pueden ser generales, intermedias e indicadores. Las generales son aquellas tan abstractas que no pueden ser directamente observadas y para su medición es necesario traducirlas a variables intermedias e indicadores. Un ejemplo de este tipo de variables es la pobreza, que requiere indicadores concretos para su medición. Por su parte, las intermedias expresan un aspecto parcial de la variable genérica como, por ejemplo, el nivel de ingresos para la medición de la pobreza.

Los indicadores representan aspectos específicos de las variables genéricas como, por ejemplo, el monto de los ingresos familiares mensuales como indicador para el nivel de ingresos. La elección de los indicadores depende de los objetivos de la investigación y pueden asumir la forma de valores absolutos, relativos (porcentajes) o ser compuestos. Por ejemplo, el número de pobres es un indicador de valor absoluto, la incidencia de pobreza (porcentaje de pobres respecto a la población total) es un indicador relativo.

2.3.2. Medición de variables

La medición de una variable consiste en el proceso de asignación de valores o categorías a las distintas características que conforman el objeto de estudio. Para que la medición se realice adecuadamente se recomienda cumplir con al menos tres requisitos básicos: la exhaustividad, la exclusividad y la precisión.

La *exhaustividad* consiste en que ninguna observación quede sin clasificarse. Por tal razón, cuando se diseña un cuestionario en el que caben

otras respuestas diferentes a las propuestas en el cuestionario, se sugiere incluir la opción de respuesta de "otros" y la categoría común "no sabe/no responde" (dirigida a aquellos que decidan no emitir ninguna respuesta).

La *exclusividad* consiste en que los distintos atributos de la variable deben ser mutuamente excluyentes, por lo que deberán definirse de manera que cualquier observación sólo pueda clasificarse en términos de un único atributo.

Ejemplo No. 4

Situación A Edad	Situación B Edad
20 y menos	Menos de 20
20 – 40	21 – 40
40 – 65	41 – 65
65 y más	Más de 66
NS/NR	NS/NR

En la situación A las personas con 20, 40 y 65 años no sabrían en qué atributo clasificarse, al estar estos valores incluidos en más de un intervalo. Mientras que en la situación B este problema desaparece cumpliéndose así el requisito de exclusividad.

En cuanto a la *precisión*, a veces es necesario realizar el mayor número de distinciones posibles, lo que contribuye a obtener información más precisa. Luego de haberse recabado la información se podrán agrupar las distintas categorías o valores de las variables. Por el contrario, no será factible desglosar los atributos después de la obtención de datos.

Ejemplo No. 5

A Nivel de instrucción	B Nivel de instrucción
Sin estudios	No sabe leer ni escribir
Primarios	Sólo sabe leer y/o escribir
Secundarios	Primarios
Superiores	Secundarios
	Educación técnica
	Educación Superior
	Otros

Si la variable "nivel de instrucción" se midiese como en la situación B se obtendrá información más precisa y detallada que si se midiese como en la situación A.

2.3.3. Escalas de medición para las variables

Se define la escala como un instrumento de medición. Por ejemplo, una regla, un termómetro, un velocímetro. En tal sentido, las operaciones matemáticas y estadísticas dependen de la escala o nivel de medición. Los cuatro niveles de medición ampliamente usados son: escala nominal, escala ordinal, escala de intervalo y escala de razón.

Escala nominal:

La medición con escala nominal implica asignar números u otros símbolos para clasificar objetos u otras observaciones. La escala nominal es un nivel mínimo de medición, los números o signos asociados a los objetos no tienen significación cuantitativa y son sólo indicaciones de presencia o ausencia de algunos atributos y características en estudio. Por ejemplo, con los números 1 y 2 se puede clasificar una población de estudiantes y no estudiantes; de usuarios y no usuarios; o de masculino y femenino. Según la regla, una escala será nominal cuando una serie incluye a todos los objetos (clasificación exhaustiva); en ningún caso un objeto puede estar en más de una categoría (son mutuamente excluyentes) y, además, las categorías se expresan en símbolos.

Ejemplo No. 6

Variable cualitativa	**Categorías**			
Asegurado en la CNS	Sí	No		
Religión	Católico	Cristiano	Bahai	Otro

Cuando se utiliza la escala nominal, la única cuantificación es contar el número de casos en cada categoría. Con esta escala sólo se puede llegar a la conclusión de cuál de las clases tiene más miembros y nada más. Es valiosa en estudios exploratorios, en los que el objetivo es hallar relaciones y no mediciones precisas; se usa frecuentemente en encuestas de opinión.

Escala ordinal:

Estas escalas tienen las características de las escalas nominales y permiten un orden de la información cuantitativa. Se aplica cuando los datos se clasifican en categorías en las que existe algún orden. Es una escala

más elaborada de medición porque nos dice que un valor dentro una categoría posee un grado mayor de la propiedad que se mide respecto a otra categoría. Sin embargo, sigue siendo débil porque no se puede hacer planteamientos numéricos sobre la diferencia entre las categorías: la ordenación establece sólo cuál categoría es mayor, mejor o preferida, pero no se refiere a cuánto mayor es o más preferida es.

Ejemplo No. 7

Variable cualitativa	**Categorías ordenadas**		
Tipo de trabajador	No calificado	Calificación básica	Calificación superior
Rango docente	Profesor A	Profesor B	Profesor A

Escala de intervalo:

Es una escala ordenada en la cual se especifica la diferencia numérica entre las mediciones y es la misma en cualquier parte de la escala. Por ejemplo, si se establece una diferencia de 5 centímetros para medir la estatura de las personas, está claro que una persona que mide 1,65 m es más alta que otra que mide 1,60 m, además, los 5 cm es la misma diferencia entre las personas que miden 1,80 m y 1,85 m; la diferencia tiene el mismo significado en cualquier parte de la escala. Otros ejemplos son el tiempo, la temperatura, la medición de la inteligencia. Esta escala se caracteriza como una unidad de medición común y constante. En este tipo de escala se puede utilizar la mayor parte de los instrumentos estadísticos.

Escala de razón:

Esta escala, además de incorporar las características de la escala de intervalo, incluye el cero real que permite calcular cocientes de mediciones. Por ejemplo, una persona que mide 1,80 m tiene el doble de la estatura de alguien que mide 90 cm.

Por lo general, las variables cuantitativas se miden en escalas de intervalo o de razón.

2.3.4. Operacionalización de las hipótesis

La *Guía para la formulación y ejecución de proyectos de investigación* del PIEB considera que "en términos simples, las hipótesis son una sospecha, de una intuición, de una posible explicación, en base a reflexiones sobre las respuestas posibles a nuestras preguntas". En tal sentido, una hipótesis es una tentativa de explicación mediante una suposición o

conjetura verosímil destinada a ser probada por los hechos (Ander-Egg 1972: 20). Las hipótesis se plantean a partir del conocimiento y reflexión teórica y se elaboran una vez que se ha formulado uno o varios problemas.

La *Guía* destaca la importancia de las hipótesis por los siguientes aspectos:

a) Son instrumentos de trabajo que permiten la coordinación en el análisis al relacionar los hechos.

b) Constituyen un hilo conductor de la investigación, dado que orientan la recolección de datos.

c) Verifican si el contenido conceptual y empírico es o no es ratificado por la realidad que se investiga.

d) Ayudan a establecer variables y a estudiar las relaciones existentes entre ellas.

Tipos de hipótesis:

Existen distintos tipos de hipótesis, como las que se orientan a la explicación y las que establecen relaciones entre variables. Las primeras pretenden demostrar o defender una posición y las segundas plantean posibles relaciones entre una o más variables.

Este último tipo de hipótesis pueden ser correlacionales y de causalidad. Las correlacionales especifican las relaciones entre dos o más variables y cómo están asociadas. Por ejemplo, la estatura de un niño de 6 años está relacionada con el tipo de alimentación. La correlación de dos variables se conoce como "correlación bivariada" y la de varias se denomina "correlación múltiple". En este tipo de hipótesis no hay relación de causalidad y no se hace referencia a variables independientes o dependientes (Hernández/Fernández/Baptista 1999: 80).

Las hipótesis de causalidad, además de incorporar relaciones entre dos o más variables, consideran la necesidad de diferenciar variables causales (independientes) que afectan una variable (dependiente); pueden ser bivariadas cuando se plantea la relación entre una variable independiente y una dependiente (por ejemplo, la pobreza depende del ingreso), y multivariadas cuando se plantea una relación de causalidad entre varias variables independientes y una dependiente (por ejemplo, la mortalidad infantil depende de la nutrición, de la atención médica y de la salud de la madre).

El proceso de operacionalización:

La operacionalización de las hipótesis es el proceso de conversión de las definiciones conceptuales mediante un conjunto de procedimientos que deben realizarse para medir las variables. Este procedimiento permite que las hipótesis puedan ser contrastadas o sometidas a la experiencia a través de su verificación con datos empíricos.

La operacionalización de las hipótesis implica considerar los siguientes aspectos:

a) Definir claramente las variables, es decir el significado de las características que se pretende investigar.

b) Establecer las escalas de medición para las variables (nominales, ordinales, de intervalo o razón).

c) Establecer el tipo de relaciones entre las variables.

d) Definir las unidades de análisis, a quién se va a investigar: familias individuos, opiniones, instituciones.

Ejemplo No. 8

Se ha propuesto una investigación que tiene como objetivo analizar los factores que determinan el nivel de ingreso de las personas ocupadas en la ciudad de La Paz. A tal efecto, se ha propuesto la siguiente hipótesis:

Los mayores niveles de ingresos de los ocupados se explican por su nivel de escolaridad y por su experiencia.

Operacionalización de la hipótesis

Variables:
La hipótesis está constituida por tres variables: ingreso de los ocupados, niveles de escolaridad y experiencia.
Definiciones:
El ingreso de los ocupados se define como el monto de dinero mensual que percibe una persona por realizar una actividad económica, ya sea como asalariado, patrón o independiente. *El independiente* es la persona que realiza actividades económicas por cuenta propia, sin relación de dependencia, con el propósito de generar un ingreso.

(Continúa en la página siguiente)

(Viene de la página anterior)

El nivel de escolaridad se define como el número de cursos aprobados en el sistema formal de educación. El sistema formal establece ocho niveles para la primaria, cuatro para la secundaria y cinco para la educación superior.
La experiencia es el número de años de trabajo que tiene una persona a lo largo de su vida. Ésta puede ser específica cuando se considera los años en un mismo tipo de ocupación, y genérica cuando no toma en cuenta el tipo de ocupación. Cuando no se cuenta con esta información se puede considerar la edad como una variable proxy de la experiencia.
Relación de variables:
El ingreso de los ocupados es la variable dependiente o explicada.
El nivel de escolaridad y la experiencia son las variables independientes o explicativa.
Por sus características, la hipótesis corresponde al tipo causal, dado que establece una relación entre las variables dependientes e independientes.
Unidad de análisis:
Está conformada por las personas ocupadas en alguna actividad económica que residen en la ciudad de La Paz[2].

2.4. El muestreo

Definidos el objetivo de la investigación, formuladas las hipótesis, determinadas las variables, su forma de medición y el tipo de investigación, se debe pasar a la fase de selección del universo de estudio y de la muestra representativa de la población.

2.4.1. Unidades de observación

Son los elementos que serán sujetos de estudio, pueden ser personas, familias, viviendas, organizaciones, espacios territoriales. La selección de las unidades de observación está en función de los objetivos de la investigación.

2 Este ejemplo servirá de guía a lo largo del texto para explicar cada una de las fases que implican la recolección de datos y el análisis de los mismos. Con fines didácticos en muchos de los posteriores ejemplos se establecerá una relación de causalidad bivariada entre el ingreso y el nivel de escolaridad, por cuanto el tratamiento de las relaciones multivariadas excede a los objetivos propuestos.

La fuente de información para los ejemplos proviene de las Encuestas de Hogares y Calidad de Vida (MECOVI) que anualmente realiza el INE en una muestra de alrededor de 5.500 familias distribuidas en las capitales de departamento (entre las que se encuentra la ciudad de La Paz), ciudades intermedias, centros poblados mayores a 2.000 habitantes y población dispersa.

Ejemplo No. 9

> Si el objetivo de la investigación es determinar los factores que inciden en el nivel de ingresos de los ocupados en la ciudad de La Paz:
>
> La unidad de observación son las personas ocupadas que residen en la ciudad de La Paz.

Definidas las unidades de observación se debe delimitar a la población (o universo de estudio) que será analizada y sobre la cual se pretende generalizar los resultados. La población "es el conjunto de unidades, para las que se desea obtener cierta información" (Sánchez Crespo 1971: 11). En la delimitación de la población se debe mencionar su ubicación en un espacio y tiempo concreto. En el Ejemplo No. 9 la población está constituida por los ocupados en la ciudad de La Paz en el año 2001.

Por razones de tiempo y costo rara vez se observa cada una de las unidades de la población (cuando se realiza esta observación se denomina censo), generalmente se observa una parte de la misma, es decir, una muestra.

2.4.2. La muestra

Se define "como una parte representativa de la población (o universo de estudio) cuyas características debe reproducir en pequeño lo más exactamente posible" (Sierra Bravo 2003: 174). La representatividad está subordinada al tamaño de la muestra y al procedimiento para la selección de las unidades muestrales.

Respecto a la relación entre la población y la muestra hay que tener en cuenta que "todo lo relacionado con poblaciones no observadas tiene que ver con las probabilidades. Sólo de lo que se observa (la muestra) se puede hablar con certeza; de lo no observado (la población) siempre hay que hablar en términos probabilísticos" (Sánchez Carrión 1995: 107). Esta relación, con algunos márgenes de error, debe permitir la posibilidad de generalizar los resultados de la muestra a la población, es decir, la validez externa de la investigación.

2.4.3. El marco muestral

Cualquiera que sea el tipo de muestreo que se adopte, previamente es necesario contar con un marco muestral que se define como el registro

o listado que comprende las unidades o elementos de la población. Es deseable que el marco contenga a todas las unidades que son de interés para el investigador y que no incluya unidades falsas o elementos ajenos porque dejaron de pertenecer a la población.

Para que el marco muestral sea un descriptor válido de la población sujeta a estudio debe cumplir al menos los siguientes requisitos:

a) El marco debe ser lo más completo posible para garantizar la representatividad de la muestra.

b) Debe estar lo más actualizado posible, con lo que se restringe la posibilidad de realizar omisiones.

c) Es necesario que cada componente de la población esté igualmente representado; se deben corregir las duplicidades: cuando éstas existen la probabilidad de ser elegido en una muestra es mayor.

d) En el marco no se deben incluir unidades que no correspondan a la población, ya sea porque dejaron de pertenecer o porque no son sujetos de estudio.

e) El marco debe ser fácil de utilizar, por ejemplo, los que se encuentran en medios magnéticos ayudan a simplificar el proceso de selección de la muestra.

Ejemplo No. 10

Dado el objetivo "determinar los factores que afectan el nivel de ingreso de las personas ocupadas en la ciudad de La Paz":

El marco muestral de acuerdo la Instituto Nacional de Estadística (INE) tiene como base la información proporcionada por el Censo Nacional de Población y Vivienda de 1992 y los datos obtenidos en la Actualización Cartográfica llevada a cabo durante los años 1998 y 2000.

Este marco está conformado por un listado de viviendas (ocupadas y desocupadas, particulares y colectivas) y número de habitantes del área urbana y rural, los cuales están clasificados mediante el orden de jerarquía de los identificadores censales: Departamento, Provincia, Ciudad, Zona Censal, Sector y Segmento Censal y Manzana en el área urbana; mientras que en

(Continúa en la página siguiente)

(Viene de la página anterior)

el área rural, los identificadores son: Departamento, Provincia, Cantón, Localidad, Sector y Segmento Censal.

Para la ciudad de La Paz la población total en el año 1992 fue de 713.378 y en el año 2001 fue de 789.293. El número de viviendas particulares en 1992 fue de 179.271 y en el 2001 fue de 220.202.

En muchos casos existen ya marcos elaborados, en especial cuando se trata de temas relacionados con la población. El INE dispone de estos marcos que son actualizados permanentemente. Cuando no se dispone de esta base de información es necesario confeccionarla. Sin embargo, existen situaciones en las cuales es imposible obtener un listado, por ejemplo, sería prácticamente imposible listar a las personas que asisten a una conferencia o espectáculo. En este caso, la solución que se puede adoptar es seleccionar una muestra por algún procedimiento aleatorio imperfecto, entrevistar a uno de cada cinco que asiste a un espectáculo (Sierra Bravo 2003: 177).

2.4.4. Tipos de muestreo

Las modalidades de muestreo se pueden agrupar en dos grandes categorías: el muestreo probabilístico y el no probabilístico. Esta categorización se realiza en función de si el azar interviene o no en las fases de selección de la muestra (Cea D'Ancona 1999: 179).

El muestreo probabilístico:

Se fundamenta en la aleatorización como criterio básico para la selección de la muestra, lo cual permite que:

a) Cada unidad de la población tenga una probabilidad igual de participar en la muestra.

b) La elección de cada unidad sea independiente de las otras.

c) Realizar el cálculo del error muestral respecto de los parámetros poblacionales.

El muestreo no probabilístico:

Incorpora criterios que no se basan en el azar para la selección de la muestra, los que están en función de un juicio personal del investigador, que puede conformar una muestra ya sea de manera *accidental*, como entrevistar a cualquier persona que pase por la calle, o de manera *intencional*, cuando tiene previo conocimiento de los elementos de la población y determina la muestra por asignación de cuotas o proporciones, en cuyo caso se pueden utilizar clasificaciones de la población por sexo, raza, religión y seleccionar miembros que sean representativos o adecuados para los fines de la investigación.

Este muestreo no requiere la existencia de un marco muestral y es apropiado para estudio exploratorios, cualitativos y para poblaciones de difícil registro como inmigrantes ilegales.

Tipos de muestreo

Probabilísticos	No probabilísticos
Simple	Por cuotas
Sistemático	Sujetos voluntarios
Estratificado	Expertos
Por conglomerados	Estratégico

Se debe aclarar que los procedimientos muestrales no son compartimientos estancos, un mismo diseño muestral puede incorporar diferente tipos de muestreo en el proceso de investigación.

En el marco del presente texto, orientado a metodologías cuantitativas, sólo se abordará los métodos probabilísticos de muestreo.

Muestreo aleatorio simple

Este tipo de muestreo constituye la referencia tanto para otros tipos de muestreo aleatorio como para el cálculo del tamaño y del error muestral.

El muestreo aleatorio simple permite elegir elementos de una población de manera que cada uno de ellos tiene la misma probabilidad de seleccionarse. La elección de las unidades de la muestra puede realizarse "sin reemplazo", cuando cada unidad de la población puede ser elegida una sola vez, y con reemplazo, cuando la unidad muestral participa en elecciones sucesivas. Lo habitual en las ciencias sociales es que el proceso de selección se realice sin reemplazo.

Para aplicar el muestreo aleatorio simple se deben seguir los siguientes pasos:

a) *Identificar cada elemento poblacional*. En base al marco muestral se asigna a cada unidad de la población un número de identificación siguiendo un orden consecutivo. Cuando el marco ya tiene números de identificación para la población, éstos pueden usarse.

b) *Seleccionar aleatoria de las unidades*. Realizada la identificación de los elementos, se procede a su selección, de manera que cada unidad tenga la misma probabilidad de ser elegida. Esta probabilidad viene determinada por la relación entre el tamaño de la muestra (n) y el total de la población (N).

Para la selección se pueden seguir dos procedimientos: el primero que consiste en registrar cada número o nombre de identificación de la población en tarjetas, las que se colocan en una urna para elegir al azar los elementos de la muestra mediante la extracción de las tarjetas. Si el muestreo es sin reemplazo la tarjeta extraída no se devuelve a la urna y si es con reemplazo se la devuelve. Este método de muestreo "con urna", aunque de fácil comprensión, no es eficiente y resulta laborioso en caso de que la población sea muy grande.

El segundo procedimiento de selección se realiza mediante la utilización de una tabla de números aleatorios. Esta tabla comprende múltiples combinaciones de números generados al azar, de forma que cada dígito presente la misma probabilidad de figurar en cualquier punto de la tabla. El procedimiento de selección de la muestra es el siguiente:

a) Se elige aleatoriamente un punto de partida que corresponda a una fila y columna y se comienza a mover en cualquier dirección, hacia arriba, hacia abajo, a la derecha o a la izquierda.

b) El número de dígitos extraídos de la tabla debe corresponder con el número de dígitos de la población del marco muestral.

c) La persona que pertenezca al número extraído pasará a formar parte de la muestra.

Ejemplo No. 11

Se requiere realizar una encuesta en una universidad de 9.660 estudiantes y se ha determinado que el tamaño de la muestra representativa es de 900 estudiantes. El marco muestral es el listado de los estudiantes de la universidad, a los cuales se les asignó un número del 0001 al 9.660.

Dado el tamaño de la muestra, la probabilidad de selección (n/N) para cada unidad de la población será: 900/9.660 = 0,093. Esto significa que de cada cien estudiantes 9,3 tienen probabilidad de estar en la muestra.

Para la extracción de la muestra se utilizará una tabla de números aleatorios (ver Anexo II). Se decide comenzar por la fila 06 columna 01 y continuar de izquierda a derecha en secuencias de cuatro dígitos, sin realizar saltos. Cuando la fila 06 concluye, se sigue con la fila 07 hasta que se alcance el tamaño de la muestra fijado. Se descartarán los números superiores a 9.660 y los números repetidos si el muestreo es sin reemplazo.

Los estudiantes seleccionados serán los que tengan los siguientes números de identificación (para este ejemplo sólo se mencionan los diez primeros) 0033, 6488, 4720, 4334, 6391, 9363, 9411, 0959, 2470 y 7054.

Paquetes estadísticos como el SPSS, STATA y otros generan automáticamente números aleatorios reduciendo sustancialmente el tiempo requerido para este tipo de selección, en especial cuando se trata de poblaciones y muestras de gran tamaño.

Entre las ventajas del muestreo aleatorio simple se destacan la facilidad de realizar cálculos estadísticos y la elevada posibilidad de inferir las características de la muestra a la población. Entre las desventajas está el laborioso proceso de listar y enumerar a las unidades de la población y la dispersión de la muestra que repercute en los costos de la investigación.

Muestreo aleatorio sistemático:

Este tipo de muestreo también requiere la existencia de un listado de la población, pero difiere del muestreo aleatorio simple en los siguientes aspectos:

a) Sólo el primer elemento de la muestra se elige aleatoriamente, ya sea mediante una tabla de números aleatorios, por sorteo o con urna, bajo la condición de que el primer número elegido sea menor al

coeficiente de elevación. El coeficiente de elevación se calcula dividiendo la población entre el tamaño de la muestra (N/n) y expresa el número de veces que la muestra está contenida en la población.

b) Calculado el coeficiente de elevación, los siguientes elementos de la muestra se obtienen sumando sucesivamente este coeficiente al primer elemento, con lo que se obtiene el segundo y a éste nuevamente se le agrega el coeficiente hasta completar el tamaño de la muestra.

Ejemplo No. 12

Si la selección de la muestra del Ejemplo No. 11 se eligiera de manera sistemática, se aplicaría el siguiente procedimiento:

Cálculo del coeficiente de elevación o intervalo de selección

N/n = 9.660/900 = 10,7 ≈ 11

Se elige al azar un número inferior al coeficiente de elevación (11), por ejemplo, el número 7 que identifica a la primera unidad de la muestra.

Las siguientes unidades se eligen sumando sistemáticamente el coeficiente de elevación empezando por el primer número seleccionado aleatoriamente. Este procedimiento da como resultado la selección de los siguientes números: 7, 18, 29, 40, 51, 62, 73, 84, ... hasta reunir como mínimo las 900 unidades que componen la muestra. Se observa que el intervalo entre cada número seleccionado es 11.

Las ventajas de este tipo de muestreo radican en que no se requiere el uso continuo de una tabla de números aleatorios u otro procedimiento de asignación aleatoria (sólo es necesario para la selección de la primera unidad). El marco muestral no debe presentar ninguna ordenación. Su principal desventaja es que se requiere un recuento constante de las unidades de la población y si el marco muestral está ordenado debe desordenarse para favorecer una selección aleatoria de la muestra.

Cuando se presentan unidades vacías como el caso de personas que no se encuentran en sus viviendas, en el muestreo simple estas unidades se pueden sustituir efectuando nuevas selecciones. En el sistemático se soluciona este problema realizando una nueva selección para todas las unidades vacías en la que el coeficiente de elevación sería el resultado de dividir la población entre el conjunto de las unidades vacías.

Muestreo aleatorio estratificado:

Este tipo de muestreo es uno de los más utilizados en la investigación social, en especial cuando se dispone de información sobre las características de la población sujeta a estudio.

El muestreo estratificado se usa por varios motivos: Aumenta el grado de precisión muestral y su eficiencia; permite trabajar o estudiar a cada estrato por separado; permite derivar estimaciones por estrato, con lo cual éstas resultan más precisas; y ayuda a resolver problemas de coordinación en el trabajo de campo.

El procedimiento del muestreo estratificado es el siguiente:

a) Se divide la población en grupos o subdivisiones denominadas estratos. La eficiencia del muestreo estratificado depende de que los grupos o estratos sean lo más homogéneos posible y de que los diferentes grupos sean lo más heterogéneos posible.

b) Cada unidad debe pertenecer a un solo grupo y ningún elemento debe quedar eliminado.

c) La conformación de los grupos o estratos se realiza en base a variables de estratificación que deben ser categóricas como, por ejemplo, el sexo, la edad, el nivel de ingreso, departamentos o regiones. Según Cochran (1981: 134), un número de variables mayor a seis genera problemas de confusión y precisión muestral.

d) Las variables de estratificación deben tener un orden, eligiendo la de mayor relevancia para conformar los primeros estratos, luego se elige la segunda variable más importante y así sucesivamente. Por ejemplo, una primera estratificación puede ser dividir la población por sexo (hombre, mujer), las segunda estratificación será dividir la población (de hombres y mujeres) por su carácter migratorio (migrantes y no migrantes).

e) Después de estratificar la población se aplica el muestreo aleatorio simple.

f) A cada una de las subdivisiones se les trata de manera independiente, aunque el método de estimación las unirá en forma global.

Ejemplo No. 13

En el ejemplo de las universidades se puede decidir estratificar la población de 9.660 estudiantes en base a los siguientes criterios:

Variables de estratificación: sexo (hombre, mujer) y tipo estudio (ciencias sociales, exactas y de la salud).

Clasificando a los estudiantes en base a las variables referidas, se encuentran seis estratos: hombres en ciencias sociales, hombres en ciencias exactas, hombres en ciencias de la salud, mujeres en ciencias sociales, mujeres en ciencias exactas y mujeres en ciencias de la salud.

	Sexo		
Tipo de estudio	**Hombre**	**Mujer**	**Totales**
Sociales	2.318	1.739	4.057
Exactas	1.449	773	2.222
Salud	2.029	1.352	3.381
Totales	**5.796**	**3.864**	**9.660**

Una aplicación concreta del muestreo estratificado es la que realiza el INE para las encuestas de calidad de vida MECOVI. A continuación se presentan los aspectos más destacables del mismo extractado de la *Documentación de la Encuesta de Hogares, Programa MECOVI 2001* (INE 2001).

Ejemplo No. 14

Encuestas de Hogares – Programa MECOVI

Universo de estudio

La investigación está dirigida al conjunto de hogares establecidos en viviendas particulares de las ciudades capitales, resto urbano y área rural de Bolivia. Excluye a las personas que habitan en viviendas colectivas.

(Continúa en la página siguiente)

(Viene de la página anterior)

Método de muestreo

Es estratificado por nivel de aglomeración de población y por región ecológica en base a una selección en dos etapas.

La población se estratifica por tipo de concentración de la población y región ecológica. Por tipo de concentración de la población las categorías son: ciudades capitales, resto urbano y área rural. Por región ecológica las categorías son : altiplano, valles y llanos.

Unidades de observación y de muestreo

La unidad de observación determinada con fines de la investigación es el hogar con todos y cada uno de los miembros que lo componen y la unidad de muestreo es la vivienda. Ésta tiene permanencia fija en el tiempo y espacio, característica que la habilita para ser utilizada como unidad de selección en el Diseño Muestral.

Procedimientos de muestreo

La primera etapa contempla la selección de una muestra de unidades primarias de muestreo (UPM´s) que en el área amanzanada (urbana) corresponden a un conjunto promedio de 110 viviendas particulares (Sector Censal); mientras que en el área dispersa (rural) las UPM´s corresponden a un conjunto de aproximadamente 40 (Segmento Censal).

La selección de las UPM´s se realiza de manera sistemática con probabilidad proporcional al tamaño de viviendas que tenga la UPM, según el marco de la Actualización Cartográfica de Población y Vivienda de 2001, dentro de cada departamento y estrato.

La segunda etapa constituye la selección de viviendas (unidades secundarias de muestreo, USM´s) de las UPM´s seleccionadas previamente; se realiza esta selección sistemáticamente con igual probabilidad mediante aplicaciones del muestreo sistemático, con arranque aleatorio.

La distribución de la muestra por estratos, se realiza a partir del total de viviendas con el fin de obtener precisión y nivel de confiabilidad requeridos.

(Continúa en la página siguiente)

(Viene de la página anterior)

La muestra de viviendas se la distribuye empleando el *Método de Asignación Potencial* en cada uno de los dominios de estudio, de la siguiente manera:

$$n_h = n \frac{N_h^a}{\sum_{h=1}^{L} N_h^a}$$

Donde: N_h Total de viviendas particulares del marco muestral en el estrato h

$N = \sum N_h$ Total de viviendas particulares del marco muestral

a Coeficiente potencial de 0,5

n Tamaño de la muestra

L Número de estratos

Clasificada la población por estratos se aplica la *afijación de la muestra*, que consiste en distribuir el tamaño de la muestra global entre los estratos, esta distribución puede ser simple, proporcional u óptima:

Simple. Se asigna el mismo tamaño de la muestra a cada estrato, con lo que se busca la misma representación de cada estrato, lo que favorece a que las estimaciones sean igualmente significativas en todos los estratos. Sin embargo, esta forma de distribución favorece a los estratos con menores unidades de población.

Proporcional. La muestra se distribuye en base al peso relativo (porcentual) de cada uno de los estratos en la población total, con lo que a los estratos con mayor número de unidades les corresponderá un mayor tamaño de la muestra global. Este criterio de afijación es el más utilizado en la investigación social. Para su cálculo se multiplica la proporción que representa el estrato en la población por el tamaño de la muestra.

Ejemplo No. 15

Para el caso de los estudiantes universitarios se requiere distribuir la muestra de 900 estudiantes en base a los criterios de afijación simple y proporcional. Los resultados serán:

			Afijación	
Tipo de estudio	**Totales**	**%**	**Simple**	**Proporcional**
Sociales	4.057	42	300	378
Exactas	2.222	23	300	207
Salud	3.381	35	300	315
Total	**9.660**	**100**	**900**	**900**

La *afijación simple* se calcula como: 900/3 = 300 que es el tamaño de la submuestra para cada estrato.

La afijación proporcional se calcula para el primer estrato de la siguiente manera: 0,42 x 900 = 378 y para los otros estratos se sigue el mismo procedimiento. Bajo este procedimiento las submuestras para cada estrato están en función del número de estudiantes por tipo de estudio.

Óptima. La muestra se distribuye añadiendo al peso relativo de cada estrato en la población la varianza del estrato. Con este criterio le corresponderá un mayor tamaño muestral a los estratos más heterogéneos. Para el cálculo de este tipo de afijación se precisa el conocimiento de la varianza poblacional, información que es difícil de conocer en los marcos muestrales a menos que se haya realizado un estudio exploratorio que incluya el valor de la varianza. Por esta razón, este tipo de afijación casi no se utiliza pese a su calificativo de óptima.

Muestreo aleatorio por conglomerados:

Los conglomerados son unidades que contienen varios elementos en su interior, por ejemplo, una manzana que contiene varias viviendas, una familia que está compuesta por varias personas, una clase que está compuesta por varios alumnos.

Este tipo de muestreo es un esquema en el cual se eligen conglomerados de elementos y se revisa completamente cada conglomerado que resulte en la muestra. En este tipo de muestreo no se requiere de la

elaboración previa de un listado de elementos, sino únicamente de los conglomerados.

Existen conglomerados de igual tamaño cuando todos ellos tienen el mismo número de elementos, y de tamaños desiguales cuando cada conglomerado contiene distinto número de elementos. Los conglomerados se eligen aleatoriamente y lo que interesa son los elementos dentro del conglomerado.

Para la *elección de los conglomerados* se deben considerar los siguientes aspectos: Los conglomerados deben estar claramente definidos y delimitados, cada unidad de la población sólo puede pertenecer a un conglomerado; el número de unidades que componen el conglomerado debe ser conocido previamente (aunque sea de manera aproximada); y los conglomerados no tienen necesariamente que estar definidos de igual manera en todos los lugares.

Entre las diferencias con el muestreo estratificado se encuentran:

a) En el muestreo estratificado se busca la homogeneidad dentro de cada estrato y la heterogeneidad entre los estratos. En el muestreo por conglomerados se busca la heterogeneidad en cada grupo y la homogeneidad entre grupos, debido a que se busca que cada conglomerado sea una representación de la variedad de los componentes del universo.

b) En el estratificado se selecciona una muestra para cada estrato, en el muestreo por conglomerados se extrae una muestra aleatoria de conglomerados.

c) En el muestreo estratificado, la unidad de muestreo es el individuo, en el de conglomerados es el conglomerado.

Muestreo polietápico por conglomerados:

Este muestreo es una extensión del muestreo por conglomerados y es muy habitual en la investigación social. En éste la unidad de muestreo final no son los conglomerados, sino las unidades que los componen. Se parte de una muestra de conglomerados, luego se extrae una nueva muestra con relación a cada uno de los conglomerados previamente elegidos.

La modalidad más habitual es la que se realiza en dos fases. En la primera se selecciona las agrupaciones de los miembros de la población de estudio, que se conocen como unidades primarias, por ejemplo, pueden

ser conglomerados de viviendas. En la segunda, se eligen aleatoriamente las unidades de la población a observar (unidades de muestreo secundarias) de los conglomerados primarios, que para el ejemplo constituyen las viviendas.

Ejemplo No. 16

En el Ejemplo No. 14, el muestreo que se aplica en las Encuestas de Hogares y de Calidad de Vida del programa MECOVI es estratificado por nivel de aglomeración de población y por región ecológica en base a una selección en dos etapas.

En el procedimiento de muestreo, *la primera etapa* contempla la selección de una muestra de unidades primarias de muestreo (UPM´s) (conglomerados) que en el área urbana corresponden a un conjunto promedio de 110 viviendas particulares, mientras que en el área dispersa (rural) las UPM´s corresponden a un conjunto de aproximadamente 40. La selección de las UPM´s se realiza de manera sistemática.

La segunda etapa constituye la selección de viviendas (unidades secundarias de muestreo, USM´s) de las UPM´s seleccionadas previamente. Se realiza esta selección sistemáticamente con igual probabilidad mediante aplicaciones del muestreo sistemático, con arranque aleatorio.

El muestreo por conglomerados se aplica generalmente en las siguientes situaciones: cuando resulta difícil elaborar una lista de todos los componentes de la población; se requiere reducir el tiempo y los costos del trabajo de campo; se realizan estudios de ámbito nacional o regional que suponen una significativa dispersión de la muestra.

Entre las ventajas de este tipo de muestreo es que no exige un listado de toda la población, sino sólo de los conglomerados, y ofrece la posibilidad de reducir costos al concentrar el trabajo en algunos puntos del muestreo. Entre sus desventajas debe mencionarse un mayor error muestral debido a la heterogeneidad de los conglomerados y la necesidad de aplicar procedimientos estadísticos complejos, en especial en el cálculo del error muestral.

2.4.5. Tamaño de la muestra

El número de unidades que se deben incluir en una muestra constituye un aspecto central en cualquier tipo de muestreo. Dependiendo de las características del diseño muestral, el cálculo del tamaño de la muestra

puede resultar una tarea compleja que requiera la participación de personas calificadas (estadísticos) para su determinación. Dadas las características de este texto se expondrán los aspectos básicos para su cálculo con sentido pragmático, a fin de ayudar al investigador en ciencias sociales que no necesariamente es un experto en técnicas estadísticas.

Para determinar el tamaño de la muestra se tienen que considerar dos factores: El nivel de confianza deseado y el error muestral permitido.

El nivel de confianza:

Es el grado de confianza o probabilidad de que los resultados que se obtengan de la muestra corresponden a la realidad. La medición del grado de confianza se basa en la distribución normal y en el teorema del límite central.

La distribución normal. Es una distribución de probabilidad que tiene forma de campana y de apariencia simétrica que muestra la probabilidad de que un evento ocurra. Una de sus características es que depende de la media y de la desviación estándar (es decir, de la dispersión de los datos respecto a la media). Transformando a una distribución normal estándar se puede medir el área debajo de la curva en unidades de desviación estándar a través de la variable Z. Los valores de la variable Z para distintos niveles de confianza se encuentran en tablas elaboradas para tal efecto, lo que permite conocer la probabilidad de acierto o error en la estimación del parámetro poblacional.

Debe aclarase que rara vez se conocen los parámetros poblacionales como la media o la desviación estándar y es a través del muestreo aleatorio que se los estima.

Tres son los niveles de confianza que más se utilizan en la investigación social y que corresponden a distintos niveles de probabilidad de que la muestra poblacional se encuentre en un determinado intervalo.

Valores "Z" estandarizados	Probabilidad de encontrarse en el intervalo
Entre + - 1 σ	0,683
+ - 2 σ	0,955
+ - 3 σ	0,997

Si se decide que la muestra tenga una probabilidad de acierto (nivel de confianza) de 95,5% en la estimación del parámetro poblacional, el valor

de Z corresponderá a dos desviaciones estándar. Esto significa que el verdadero parámetro poblacional se encuentra entre más y menos dos desviaciones estándar. Si se opta por un nivel de confianza de 99,7% el parámetro poblacional se encontrará entre más-menos tres desviaciones estándar. El nivel de confianza más utilizado es de 95,5%.

El teorema del límite central. Dice que independientemente de la distribución de la población, sea ésta normal o no, la distribución de una muestra lo suficientemente grande se aproxima a una distribución normal.

Berenson y Levine plantean: "¿Qué tamaño de muestra es lo suficientemente grande? Se ha realizado una gran investigación estadística sobre este tema. Como regla general, los estadísticos han encontrado que para la mayor parte de las distribuciones poblacionales siempre que el tamaño de la muestra sea por lo menos de 30, la distribución muestral de la media será aproximadamente normal" (1992: 22).

El teorema del límite central es de crucial importancia para la inferencia estadística ya que permite llegar a conclusiones sobre una población sin tener que conocer la distribución de la población.

El error muestral:

Como la muestra representa una parte de la población y se pueden obtener n muestras de una misma población, siempre existirán diferencias entre los resultados obtenidos de la muestra respecto a la población. Estos resultados se denominan estimaciones o estadísticos de los parámetros poblacionales. Bajo esta consideración se define al error muestral como la diferencia entre las estimaciones (estadísticos) muestrales y los parámetros poblacionales. Por ejemplo, la diferencia entre la media poblacional y la media muestral.

Para el cálculo de error muestral intervienen los siguientes aspectos: El nivel de confianza deseado; la desviación estándar; el tamaño de la muestra; y el tipo de muestreo adoptado.

El nivel de confianza afecta el error muestral en cuanto un mayor nivel de confianza implica que la probabilidad de acierto se encuentre en un intervalo mayor. Por ejemplo, si se fija un nivel del confianza de 99,7%, implica que la media o la proporción de la población se encontrará en un intervalo comprendido entre más-menos tres desviaciones estándar, si se reduce el nivel de confianza a 95,5% se estima que la media o proporción poblacional se encontrará entre más-menos dos desviaciones estándar. Por lo tanto, un mayor nivel de confianza implica un menor error muestral.

La desviación estándar que determina el grado heterogeneidad o dispersión de una población también afecta al error muestral. Cuando las muestras son más heterogéneas existe mayor probabilidad de error muestral. Generalmente, la desviación estándar de la población no se conoce. En algunos casos se la puede estimar de estudios o información anterior y, en otras ocasiones, se puede desarrollar un estudio exploratorio o una conjetura sobre este parámetro poblacional.

2.4.6. Sobre las proporciones

Las proporciones poblacionales (que pueden ser expresadas en porcentajes multiplicando la proporción por cien) se usan en general para variables categóricas o cualitativas (que utilizan escalas nominales u ordinales) y son muy empleadas en las ciencias sociales. En muchos casos se trata de estimar estas proporciones a partir de una muestra. Por ejemplo, se puede tratar de estimar el porcentaje o proporción de niños en una determinada población, el porcentaje de ocupados, el porcentaje de viviendas de tres cuartos en una ciudad.

Las proporciones (P) o probabilidades se calculan dividiendo el número de casos deseados o exitosos entre el tamaño de la población. Por ejemplo, el número de niños entre la población total. Asimismo, se puede calcular la proporción o probabilidad de casos no deseados o no exitosos (Q), considerando que los casos no exitosos son un complemento de los casos de éxito, con la siguiente fórmula: Q = (1-P)

Para la media de las proporciones y la desviación estándar se aplica las siguientes fórmulas:

Media = P Donde P es la proporción de éxito

Desviación estándar $S = \sqrt{P \times Q}$ Donde Q es la proporción de fracaso

Ejemplo:

En una población de 9.660 estudiantes se determina que 60% de los estudiantes tiene notas por encima de 50 puntos.

En este caso la media es el 60% de estudiantes con notas mayores a 50 puntos.

(Continúa en la página siguiente)

(Viene de la página anterior)

La desviación estándar será:

$$S = \sqrt{P \times Q} = \sqrt{40 \times 40} = 49$$

Cuando se trata de proporciones (variables cualitativas o categóricas) y no se cuenta con información previa sobre la desviación estándar de la población, ésta se puede calcular considerando una proporción de éxitos (P) del 50% y la de fracasos (Q) de 50%.

El tamaño de la muestra, incide de manera importante en el error muestra. En general, a medida que crece el tamaño de la muestra, decrece el error muestral, cualquiera sea el valor de la desviación estándar de la población.

El tipo de muestreo adoptado afecta el error muestral por el procedimiento de selección de las unidades de la muestra. Sobre los distintos tipos de muestreo aleatorio debe considerarse:

a) Que el muestreo sistemático tiene errores de muestreo equivalentes al del muestreo simple.

b) Que el muestreo estratificado presenta menores errores que el muestreo simple del mismo tamaño, siempre y cuando exista heterogeneidad entre los estratos.

c) Que el muestreo por conglomerados genera mayores errores muestrales, sobre todo cuando mayor es el tamaño de los conglomerados (Cea D'Ancona 1999: 174).

Cálculo del error muestral:

El error muestral en términos operativos no es otra cosa que la desviación típica de la muestra, la fórmula de cálculo varia según se trate de poblaciones infinitas o finitas, de variables cuantitativas medidas en escalas de intervalo o de razón o de variables cualitativas o categóricas medidas en escalas nominales u ordinales que se expresan en proporciones. Las fórmulas se ajustan cuando se trata de muestreo estratificado. Las fórmulas que se presentan suponen que el muestreo se realiza sin reemplazo.

Fórmulas del error muestral para muestras simples y sistemáticas:

	Población infinita	**Población finita (menos de 100.000)**
Error típico de la media	$e_{\bar{x}} = \sqrt{\frac{S^2}{n}}$	$e_{\bar{x}} = \sqrt{\frac{S^2}{n}\frac{N-n}{N-1}}$
Error típico de la proporción	$e_p = \sqrt{\frac{P \times Q}{n}}$	$e_p = \sqrt{\frac{P \times Q}{n}\frac{N-n}{N-1}}$

Donde: S^2 es la desviación estándar elevada al cuadrado, es decir, la varianza
N es la población total es el tamaño de la muestra
n es la proporción de la muestra que posee el atributo en cuestión
Q es la proporción que no posee el atributo

Fórmulas del error muestral en muestras estratificadas:

Error típico de la media	Error típico de la proporción
$e_{\bar{x}} = \sqrt{\frac{\sum_{i=1}^{n} n_i S_i^2}{n^2}}$	$e_{px} = \sqrt{\frac{\sum_{i=1}^{n} n_i P_i Q_i}{n^2}}$

Donde: n_i es el tamaño de la muestra del estrato i
S_i^2 es la varianza (desviación estándar al cuadrado) de la variable de interés de la población del estrato i
P_i es la proporción de la muestra en el estrato i que posee el atributo
Q_i es la proporción de la muestra que no posee el atributo

Ejemplo No. 17

Consideremos el caso de la universidad donde estudian 9.660 personas, población de la que se extrajo una muestra de 900 estudiantes. En la muestra se observa que 63% de los estudiantes tienen notas mayores a 50 puntos. Se quiere estimar el error muestral y a partir de él inferir cuál será el parámetro poblacional con un nivel de confianza del 95,5% (es decir, en un intervalo de más-menos 2s).

(Continúa en la página siguiente)

(Viene de la página anterior)

En este caso P = 63% y Q = (100 – 63) = 37% (Expresado en porcentajes)

Aplicando la fórmula del error muestral para proporciones de una población finita con muestreo simple, se tiene:

$$e_p = \sqrt{\frac{P \text{x} Q}{n} \frac{N - n}{N - 1}} = e_p = \sqrt{\frac{63 \text{x} 37}{900} \frac{9.660 - 900}{9.660 - 1}} = 1{,}53\%$$

(Error muestral)

El límite de confianza se calcula con la siguiente formula:

$P \pm Z.e_p$, donde P es la proporción de la muestra, e_p el error muestral y Z nivel de confianza previamente elegido medido en unidades de desviación estándar.

Los límites de confianza serán: el superior 63 + (2 x 1,53) = 66,06 y el inferior 63 - (2 x 1,53) = 59,94. Estos resultados significan que habrá 95,5% de probabilidad de que el porcentaje de estudiantes con notas superiores a 50 puntos se encuentren entre 66,06% y el 59,94% de la población estudiantil.

Mediante los límites de confianza se establece la probabilidad de que en el intervalo, dado un nivel de confianza, se encuentre la verdadera proporción de la población.

Cálculo de tamaño de la muestra:

Una vez determinado el nivel de confianza o el grado de probabilidad de que la muestra refleje los verdaderos parámetros poblacionales y el nivel de error o diferencias que se puede aceptar entre los estadísticos muestrales y las medidas de la población, se puede calcular el tamaño de la muestra para poblaciones infinitas y finitas para los casos de variables que se expresen como valores numéricos en escalas de intervalo o que se expresen como proporciones (porcentajes), asumiendo que el muestreo se realiza sin reemplazo.

Fórmulas para calcular el tamaño de la muestra (muestreo simple y sistemático):

Para muestras	Población infinita	Población finita (menos de 100.000)
De intervalo	$n = \frac{Z^2 \text{x} S^2}{e^2}$	$n = \frac{Z^2 \text{x} S^2 \text{x} N}{e^2(N-1) + Z^2 S^2}$
De proporciones	$n = \frac{Z^2 \text{x} P \text{x} Q}{e^2}$	$n = \frac{Z^2 \text{x} P \text{x} Q \text{x} N}{e^2(N-1) + Z^{2.} P \text{x} Q}$

Donde: Z representa el nivel de confianza previamente elegido y se mide en unidades de desviación estándar

S^2 es la varianza de la población (desviación estándar al cuadrado)

e es error de muestreo admitido por el investigador, el cual se establece *a priori*

Tamaño de la muestra en el muestreo estratificado:

En el muestreo estratificado, el tamaño (n) de la muestra es la suma de los tamaños de las muestras de cada estrato. Cuando se tiene el tamaño de la muestra de toda la población, ésta se puede distribuir asignando una ponderación (afijación proporcional) para cada estrato Nh/N. Donde Nh es la población del estrato h. Con este criterio, a los estratos más grandes se les asigna un tamaño mayor de muestra y a los más pequeños un tamaño menor. Este método es ampliamente usado en la práctica. La fórmula para distribuir el tamaño de la muestra es:

$$n_h = (\frac{N_h}{N})\, n$$ donde h representa el número de estrato

Error de muestreo admitido. Considerando que el error muestral interviene en la decisión sobre el tamaño de la muestra y éste se determina *a priori*, se debe considerar que un mayor tamaño de la muestra implica una mejor estimación de los parámetros poblaciones y, por consiguiente, una reducción del error muestral.

Sin embargo, una muestra mayor tiene efectos sobre los costos y tiempo al momento de recolectar la información de la población seleccionada.

Por lo que se deberá encontrar una situación intermedia entre el error y tamaño de la muestra. Los errores comprendidos entre el 2,5% y el 2% son los más frecuentes en la investigación social. En el Anexo III se incorpora una tabla para poblaciones infinitas (más de 100.000 habitantes), en las que se establecen tamaños de muestra con un nivel de confianza de 95,5% para distintos valores de las proporciones P y Q.

Ejemplo No. 18

Tomando como referencia los datos del ejemplo de la universidad, se propone realizar un estudio sociológico general sobre esta población. Por tal razón, es necesario contar con una muestra que permita trabajar con un nivel de confianza del 95,5% y con un margen de error de 3% (la población estudiantil es de 9.660 personas).

Dado que la población es finita y las variables para el análisis sociológico son en su mayor parte cualitativas, el tamaño de la muestra que cumpla con los requisitos especificados será:

$$n = \frac{Z^2 \text{x} P \text{x} Q \text{x} N}{e^2(N-1) + Z^2 \text{x} P \text{x} Q} \Rightarrow n = \frac{2^2 \text{x} 50 \text{x} 50 \text{x} 9.660}{3^2(9.660-1) + 2^2 \text{x} 50 \text{x} 50} = 997 \text{ estudiantes}$$

Como no se conoce la proporción de P, se supone el caso más común de P = 50

Como alternativa se puede considera que P = 65; en tal caso la muestra será:

$$n = \frac{2^2 \text{x} 65 \text{x} 35 \text{x} 9.660}{3^2(9.660-1) + 2^2 \text{x} 50 \text{x} 50} = 907$$

Si se requiere calcular la muestra para los estratos de estudiantes por tipo de disciplina, se procederá distribuyendo la muestra de acuerdo al peso o ponderación de cada estrato en la población total y aplicando la fórmula,

$n_h = (\frac{N_h}{N}) n$ con la cual se obtienen los siguientes tamaños de muestra:

Tipo de estudio	Totales	%	Tamaño de muestra
Sociales	4.057	42	419
Exactas	2.222	23	229
Salud	3.381	35	349
Total	**9.660**	**100**	**997**

La determinación del tamaño de la muestra cuando se trata de encuestas de carácter nacional como la Encuesta de Hogares y Calidad de Vida del Programa MECOVI o la Encuesta Nacional de Salud (ENDSA), que adoptan diseños muestrales combinados de estratificación y por conglomerados, requiere fórmulas más complejas.

Ejemplo No. 19

En el caso de la Encuesta de Hogares y de Calidad de Vida del Programa MECOVI, el muestreo, como se mencionó, "es estratificado por nivel de aglomeración de población y por región ecológica en base a una selección en dos etapas".

El tamaño de muestra ha sido calculado en base a los indicadores de la tasa de incidencia de pobreza (37,4%), gastos en consumo de promedio, más un ajuste de múltiplo y tasa de cobertura. Se ha considerado un nivel de confiabilidad de 95% y un error relativo permisible máximo de 5% en las estimaciones.

Para el tamaño de la muestra la fórmula aplicada fue la siguiente:

$$n = \frac{\left(\sum N_h^a\right)\left(\sum N_h^{2-a} S_h^2\right)}{\frac{N^2 \varepsilon^2}{Z^2_{1-\beta/2}} + \sum N_h \, S_h^2}$$

Donde:

N_h	Total de viviendas particulares del marco muestral en el estrato h
S_h^2	Varianza de la variable principal de estudio X, en el estrato h
ε	Error permisible global
$Z^2_{1-\beta/2}$	Valor tipificado de la distribución normal a β de nivel de significación
$n = \sum N_h$	Total de viviendas particulares del marco muestral
a	Coeficiente potencial de 0,5

La distribución de la muestra por dominios de estudio y estratos se determinó a partir del total de viviendas con el fin de obtener cierta precisión y nivel de confiabilidad requeridos.

(Continúa en la página siguiente)

(Viene de la página anterior)

Se distribuye la muestra de viviendas empleando el Método de Asignación Potencial en cada uno de los dominios de estudio, de la siguiente manera:

$$n_h = n \frac{N_h^a}{\sum_{h=1}^{L} N_h^a}$$

Como resultado, el tamaño de la muestra de viviendas fue de 5.744 viviendas y 652 Unidades Primarias de Muestreo (UPM's).

La distribución de la muestra por departamentos en el 2001 tuvo estos resultados:

Divisiones	Viviendas
Total	**5.744**
Chuquisaca	420
La Paz	1.440
Cochabamba	1.036
Oruro	428
Potosí	572
Tarija	400
Santa Cruz	1.004
Beni	344
Pando	100

2.4.7. Resumen de las etapas del muestreo

a) Seleccionar las unidades de observación: Consiste en definir las unidades o sujetos de una población que deberán incorporarse en el proceso del muestreo.

b) Identificar el marco muestral: Significa conocer la población a la cual se aplicará la muestra, por ejemplo, el censo de población sirve de marco muestral para diferentes muestras aplicadas a las ciencias sociales.

c) Elegir el procedimiento: Que pueden ser aleatorios, determinísticos y estratificados.

d) Estimar los errores muestrales: Consiste en evaluar las diferencias entre los valores medios estadísticos de la muestra y la población.

e) Determinar el tamaño: Consiste en definir el número de sujetos que se incorporará en la muestra con carácter representativo.

3. Recolección de datos

Seleccionadas las unidades sujetas a investigación y definido el tamaño de la muestra, la siguiente fase constituye la recopilación de dados. Si la investigación tiene carácter cualitativo, existen diferentes técnicas para la recolección de la información como las distintas formas de observación: experimental, estructurada, participante, entre otras. En la investigación que privilegia los aspectos cuantitativos, la encuesta es la forma que por excelencia se utiliza para la recopilación de la información (Barragán 2003: 204).

3.1. La encuesta

Se define como un procedimiento estandarizado para recabar información (oral o escrita) de una muestra de sujetos (Cea D'Ancona 1999: 239). Entre sus principales características se destacan:

a) La información se obtiene mediante respuestas de los sujetos encuestados, por lo que cabe la posibilidad de que la información no necesariamente refleje la realidad del tema que se investiga.

b) Se puede incorporar una amplia gama de temas. Puede incluir aspectos objetivos (datos) y subjetivos (opiniones y actitudes).

c) La información se recoge de forma estructurada. Se formulan las mismas preguntas en el mismo orden y a cada uno de los individuos de la muestra, lo que permite la comparación de los sujetos.

d) Las respuestas se agrupan y cuantifican para examinarlas posteriormente mediante técnicas estadísticas.

e) La confiabilidad de la información dependerá de la existencia de errores de muestreo y de errores ajenos al muestreo (diseño de cuestionario, el trabajo de campo y el procesamiento de datos).

3.1.1. Ventajas y desventajas de la encuesta

Entre las ventajas se pueden mencionar las siguientes:

a) Obtener información diversa de un amplio conjunto de personas que pueden estar ubicadas en distintas áreas geográficas.

b) Elevado grado de fiabilidad que se logra gracias a la estandarización de las respuestas y por la formulación de las preguntas.

c) Permite la comparación de datos entre periodos, áreas geográficas y grupos de personas.

d) Permite la aplicación de la teoría de la probabilidad y del muestreo y de la significación estadística.

Las principales desventajas son:

a) No es aconsejable aplicarla a poblaciones con dificultades en la comunicación (niños pequeños, analfabetos).

b) La información que se obtiene se halla condicionada por la formulación de las preguntas y por la veracidad de las respuestas.

c) No permite conocer a profundidad la realidad que se investiga dado que la mayor parte del cuestionario se estructura con preguntas cerradas, por lo que es recomendable complementarla con otras estrategias de investigación.

d) La presencia del entrevistador puede provocar reacciones no favorables para la emisión de las respuestas.

e) La realización de una encuesta requiere de la organización de un trabajo de campo muchas veces costoso y complejo que puede convertirse en una restricción de no contarse con adecuados recursos económicos.

3.1.2. Errores en la encuesta

En el proceso de elaboración y aplicación de la encuesta pueden ocurrir errores tales como:

Errores de cobertura. Son aquellos que se presentan cuando no se cubre en su totalidad las unidades incluidas en la muestra.

Errores de no respuesta. Se presentan cuando algunas personas no participan (por su no localización, negativa personal u otra causa) o sólo responden algunas preguntas del cuestionario. La importancia de este error dependerá de la magnitud de la "no respuesta".

Errores en el diseño del cuestionario. Provienen de la formulación y el orden de las preguntas y del modo de administrarlo (entrevista personal, por teléfono o correo).

Errores atribuibles al entrevistador. Son los que se producen cuando se cambia el sentido de las preguntas o no se siguen las instrucciones para la aplicación del cuestionario.

Errores atribuibles al entrevistado. Se presentan cuando las respuestas no son veraces debido a fallas de memoria, factores de situación social como la tendencia a ajustar las respuestas a lo deseable socialmente.

3.1.3. Fases para la realización de la encuesta

El procedimiento para la encuesta incluye las siguientes fases:

a) Diseño del cuestionario.

b) Selección de la modalidad de encuesta.

c) Proceso de entrevista (trabajo de campo).

d) Codificación y tabulación.

3.2. El cuestionario

En la investigación cuantitativa mediante encuesta el instrumento básico para la recolección de información es el cuestionario estandarizado. La finalidad del cuestionario es obtener de manera sistemática y ordenada información sobre las variables de una muestra o población objeto de la investigación. El cuestionario no es una técnica exclusiva para la encuesta; puede aplicarse a otras estrategias de investigación como el experimento.

El cuestionario consiste en un listado de preguntas estandarizadas y su formulación es idéntica para cada encuestado. Fowler (1988) recomienda que antes de diseñar un cuestionario se ponga por escrito el propósito de la encuesta y se elabore un listado de las variables independientes, dependientes y las de control que surgen de las hipótesis a contrastar.

3.2.1. Las preguntas

El elemento básico del cuestionario son las preguntas, es por ello que su validez depende de la clase de preguntas y de su adecuada formulación. Desde el punto de vista de la investigación social, las preguntas son la expresión en forma interrogativa de las variables empíricas o indicadores respecto a los cuales interesa obtener información (Sierra Bravo 2003: 307).

Es fundamental en el cuestionario que las preguntas reúnan las condiciones de exhaustividad, exclusividad y precisión con el propósito de medir posteriormente las variables.

Elaboración de preguntas. Para realizar esta actividad se deben tomar en cuenta los siguientes criterios:

a) Formular preguntas relevantes para la consecución de los objetivos.

b) Preguntas breves y fáciles de comprender, lo que implica previamente conocer el nivel educativo y el vocabulario de la población.

c) Evitar palabras ambiguas que dificulten la interpretación de las preguntas como de las respuestas. Palabras como "demasiado", "mayor", "poco", "joven", "usualmente" tienen esta característica.

d) No emplear palabras que generen reacciones, lo que disminuye el grado de respuestas. Palabras como "cholo", "racista", "drogadicto" generan respuestas falsas o no respuesta. Asimismo, se recomienda no utilizar frases hipotéticas como "¿Qué haría si...?" o "¿Le gustaría...?", o preguntas que inevitablemente producen respuestas favorables como, por ejemplo, "¿Le gustaría tener mayores ingresos?". Las preguntas deben formularse de manera objetiva (neutra) a fin de no influir en la respuesta.

e) No redactar preguntas en forma negativa. Las formulaciones negativas se comprenden peor que la formulación positiva, en especial cuando se pide que se manifieste el grado de acuerdo o desacuerdo ante determinadas cuestiones. Por ejemplo, la pregunta "¿No deberían trabajar los niños?" se puede formular de forma no negativa: "¿Los niños deberían trabajar?".

f) Las preguntas no deben referirse a varias cuestiones al mismo tiempo. Preguntas como "¿Las últimas elecciones fueron transparentes y con

amplia participación ciudadana?" no facilitan la interpretación correcta y puede que la respuesta a la primera parte sea afirmativa, pero la correspondiente a la segunda sea negativa. Este criterio también es válido para las preguntas cerradas, las que no deben contener dos o más ideas afines en una misma categoría.

g) Evitar preguntas que requieran cálculos mentales o que impliquen recurrir a la memoria. Estas preguntas pueden afectar la fiabilidad de las respuestas o puede que no se las responda. Se sugiere preguntar por acontecimientos del pasado inmediato (última semana, o el último mes). Tomar como referencia acontecimientos o fechas importantes, por ejemplo, "¿Desde año nuevo ha realizado algún viaje?" o proporcionar al entrevistado una o más señales de memoria.

3.2.2. Tipos de preguntas

Existe una variedad de clasificaciones en función a distintos criterios. La más utilizada es según la contestación que admite el entrevistado, en tal sentido, las preguntas pueden ser cerradas y abiertas.

Preguntas cerradas:

También denominadas precodificadas o de respuesta fija, son aquéllas en las que las alternativas de respuesta están predeterminadas. Para facilitar el proceso informático se le asigna un código numérico. En la codificación de las respuestas deberá seguirse las mismas pautas en todas las preguntas del cuestionario. Por ejemplo, codificar siempre las respuestas SÍ como 1 y las NO como 2, y utilizar el 0 si no contesta.

En el proceso de diseño estas preguntas deben cumplir con los requisitos de exhaustividad, precisión y exclusión. Si en el cuestionario no se agotan todas las posibilidades, se deberá incluir la opción de "otros".

Las preguntas cerradas pueden ser uniopcionales o de opción múltiple. Las primeras sólo aceptan una respuesta de un conjunto de atributos; la múltiple acepta una o más opciones, en este caso se deben dar instrucciones expresas sobre cuántas alternativas de respuesta se admiten.

Ejemplo No. 20

Preguntas cerradas de tipo uniopcional	
¿Cuál es su estado civil o conyugal actual? 1. Soltero/a 2. Casado/a 3. Conviviente o concubino/a 4. Separado/a 5. Divorciado/a 6. Viudo/a	*¿A cuál religión o culto pertenece actualmente?* 1. Ninguna 2. Católica 3. Protestante/evangélica 4. Otra religión o culto de origen cristiano (especifique) ....

Fuente: INE 2001.

Los números a la izquierda de cada pregunta son los códigos asignados.

Ejemplo No. 21

Preguntas cerradas de tipo multiopcional	
¿Qué idiomas habla? 1. Quechua 2. Aymará 3. Castellano 4. Guaraní 5. Extranjero 6. No puede hablar 7. Otro nativo (especifique)	*Las rentas que recibe son por concepto de:* 1. Intereses 2. Alquileres 3. Jubilación 4. Benemérito 5. Invalidez 6. Viudez 7. Otro (especifique) ....

Fuente: INE 2001.

Las preguntas cerradas tienen algunas ventajas como:

a) La rapidez y comodidad de su registro.

b) La posibilidad de focalizar las respuestas a los aspectos relevantes del objeto de la investigación.

c) La reducción de tiempo para el procesamiento de las respuestas.

d) La comparación de las respuestas.

Sin embargo también conllevan inconvenientes:

a) Coartan las opciones de respuesta.

b) El diseño de las preguntas cerradas es laborioso, exige el previo conocimiento de la realidad que se investiga. No sólo implica cómo formular la pregunta, sino también definir qué categorías de respuesta se considerará, los niveles de medición (nominal, ordinal, de intervalo, de razón) y qué códigos se asignará a cada respuesta.

c) La duda de si las distintas opciones de respuesta son igualmente interpretadas por todos los encuestados. A una misma respuesta los encuestados pueden atribuirle significados diferentes. Respuestas como "mucho", "poco" provocan diferentes interpretaciones: mientras que para algunos comer tres panes al día puede parecer poco, para otros puede resultar mucho.

Preguntas abiertas:

Son aquellas que no establecen respuestas predeterminadas y los sujetos tienen la libertad de contestar sin limitarse a escoger una serie de alternativas. Estas preguntas pueden ser de carácter cuantitativo o cualitativo; el primer caso implica respuestas de contenido numérico y, el segundo, una respuesta que puede ser de opinión, deseo o calificación.

Ejemplo No. 22

Preguntas abiertas cuantitativas
¿Cuántos años cumplidos tiene usted? .....
¿Cuál es su salario mensual? .....
¿Cuántos miembros tiene su familia? .....
¿Cuántas veces fue al médico en los últimos tres meses? .....
Preguntas abiertas cualitativas
¿Qué opina usted del funcionamiento del Caja Nacional de Salud?.....
¿Qué cambios deberían realizarse en la Reforma Educativa?.....
¿Dónde le gustaría vivir?.....

El entrevistador tiene que anotar literalmente la respuesta del encuestado, no debe introducir ninguna modificación que pudiera alterar su significado. Este requisito puede requerir mayor tiempo y la posibilidad de cometer un mayor número de errores.

Las preguntas abiertas son útiles en estudios exploratorios y cuando no se dispone de un conocimiento previo suficiente del tema que se investiga o cuando no se prevén todas las posibles respuestas.

Las preguntas abiertas tienen como principal inconveniente la traducción de las respuestas en categorías que las resuman, actividad que se conoce como el "cierre de preguntas abiertas", la cual se efectúa mediante la extracción de una muestra de los cuestionarios llenados para proceder a la codificación de las respuestas abiertas. Este procedimiento implica transcribir las respuestas emitidas, buscar en las respuestas términos comunes y, en función de ello, agruparlas en un número reducido de categorías y codificarlas. Cada categoría debe cumplir los requisitos de las preguntas cerradas (exhaustividad, exclusividad y precisión).

En general, las preguntas abiertas presentan las siguientes ventajas:

a) Proporcionan mayor información de los aspectos que se investiga.

b) Su formulación es más sencilla.

c) Ocupan menor espacio en el cuestionario.

d) El encuestado se puede expresar en sus propias palabras.

Preguntas filtro:

Son aquellas que se formulan con anterioridad a otra con la finalidad de eliminar a los sujetos a los que no procede realizar la pregunta siguiente. En estas preguntas se debe incluir la instrucción inmediatamente después de la respuesta.

Ejemplo No. 23

¿Durante este año se matriculó en algún nivel de educación escolar o superior?
Sí ..... (pase a la siguiente pregunta) No ..... **¿A qué nivel de educación se inscribió este año?** 1. Primaria 2. Secundaria 3. Normal 4. Universidad 5. Técnico 6. Otros (especifique) .....

Fuente: INE 2001.

Preguntas de ponderación:

Son preguntas abiertas donde se pide una valoración sobre una situación concreta.

Ejemplo No. 24

En una escala de 0 a 5 califique el desempeño del comité de vigilancia
.....

3.2.3. Orden de las preguntas

Constituye una parte relevante en el diseño del cuestionario. La calidad de las respuestas puede afectarse no sólo por la redacción de las preguntas, sino también por su ubicación en el cuestionario. Sobre el ordenamiento de las preguntas existen algunos criterios que se señalan a continuación:

a) El cuestionario debe tener un espacio para incorporar los datos de control como el número de orden del cuestionario, nombre, dirección y teléfono del entrevistado; fecha, hora de la entrevista y cualquier otro dato que se estime relevante para el control del trabajo de campo.

b) Las preguntas iniciales deben ser generales y despertar el interés del encuestado, para luego ir introduciendo preguntas más complejas (técnica del embudo); en algunos casos se puede empezar por las preguntas más específicas, para pasar luego a las preguntas más generales (embudo invertido). Este procedimiento se conoce como el método encauzado (Namakforoosh 2000: 183).

c) Las preguntas claves deberían ubicarse en el centro del cuestionario, para realizarlas una vez que se haya despertado el interés del entrevistado y se haya generado un ambiente favorable.

d) Agrupar las preguntas sobre una misma temática.

e) Diseñar un cuestionario que no sea extenso con el propósito de no fatigar al encuestado para que no mermen la calidad de sus respuestas. Cabe aclarar que no es el número de preguntas lo que determina la amplitud de un cuestionario, sino la duración media de la entrevista. Se aconseja que la encuesta no sobrepase la hora.

3.2.4. Formato del cuestionario

El cuestionario debe diseñarse de forma atractiva y fácil de responder; las recomendaciones para lograr este propósito son las siguientes:

a) Numerar las preguntas y capitular los temas cuando éstos se presenten.

b) Espaciar las preguntas; a veces por reducir el espacio se provoca efectos no deseados como una menor cooperación y mayor probabilidad de errores.

c) En las preguntas abiertas proporcionar espacio suficiente para anotar las respuestas.

d) Evitar que las preguntas queden partidas entre páginas.

e) Una pregunta larga que incluye varias partes no debería estar seguida de una pregunta breve al final de página.

f) Proporcionar instrucciones al entrevistador, en lugares apropiados y de fácil identificación, mediante la impresión en otro tipo de letras (cursivas, mayúsculas). Las instrucciones se colocan delante de la pregunta para aclarar la forma de realizar o responder la pregunta.

Las que se refieren al modo de registrar las respuestas se colocan al final de cada pregunta cuando sea necesario.

3.2.5. Evaluación y prueba del cuestionario

Tanto el diseño como las consecuencias de las preguntas deben evaluarse. Es importante que el cuestionario se examine antes de aplicarlo. Para asegurar su entendimiento se debe tomar una muestra pequeña de la población en estudio, aplicarla y medir el grado de comprensión y posible modificación del cuestionario. Este proceso se conoce como prueba piloto o pre-test, y permite evaluar la adecuación del cuestionario en los siguientes aspectos:

a) Si las preguntas tienen sentido y se comprenden.

b) Si la categorización de las respuestas y su codificación es correcta.

c) Si las instrucciones que figuran en el cuestionario se entienden, así como el formato de las preguntas filtro.

De los resultados de la evaluación surgirá la necesidad de revisar algunas preguntas, la secuencia de los temas, la inclusión de nuevas preguntas y el porcentaje estimado de "no respuesta".

La prueba piloto proporciona información sobre la validez del marco muestral, la variabilidad de la población, preparación de los entrevistadores y el costo aproximado del trabajo de campo.

3.3. Modalidades de encuesta

Existen tres modalidades de encuesta en función de cómo se administra el cuestionario: entrevista personal, telefónica y por correo. Por ser la más generalizada se abordará la entrevista personal también denominada "cara a cara".

3.3.1. Entrevista personal

Consiste en la administración del cuestionario mediante entrevista personal a cada uno de los individuos seleccionados en la muestra. Las entrevistas pueden llevarse a cabo en el domicilio de las personas, en su lugar de trabajo o en lugares donde se pueden encontrar a los sujetos a entrevistarse.

Para un buen resultado de las entrevistas es necesario contar con personas previamente capacitadas con el propósito de la investigación, habilidad para realizar preguntas y para ubicar los sitios donde se aplicará la encuesta, responsabilidad y ética profesional.

Por lo general, en el proceso de la entrevista intervienen tres personas: investigador, entrevistador y entrevistado. El investigador encarga a alguien la realización de la entrevista, pero debe mantener el control y la supervisión de ésta. El investigador puede crear varias técnicas de control como incorporar una sección final en el cuestionario donde se anoten la dirección, teléfono y nombre del entrevistado, datos que podrán servir para la validación de las respuestas, procedimiento que se realiza mediante la selección de una muestra de los cuestionarios los cuales nuevamente se aplican al entrevistado en forma completa o seleccionando algunas preguntas (Namakforoosh 2000: 143). Otra técnica de control o de validación es crear dos equipos de entrevistadores, un equipo grande para realizar las entrevistas y otro equipo pequeño para verificar la aplicación de las entrevistas.

Es importante cuidar la forma en que los entrevistadores seleccionan a los entrevistados, ya que se puede alterar el diseño del muestreo. Asimismo, se debe especificar los tiempos de la entrevista; en muchos casos los entrevistadores pueden realizarla de manera muy rápida con posteriores efectos sobre la validez de los resultados.

El entrevistador debe estar adiestrado y adecuadamente preparado para:

a) Formular correctamente las preguntas del cuestionario.

b) Asegurar la adecuación de las respuestas y su correspondiente anotación.

c) Tomar decisiones en el campo sin la asistencia del supervisor.

d) Comprobar la comprensión de las preguntas y la consistencia de las respuestas.

e) Recoger información complementaria sobre el entorno y características personales del entrevistado.

Para que la comunicación entre el entrevistador y el entrevistado sea fácil y efectiva, los entrevistadores deben ser relativamente similares al entrevistado (misma manera de vestir, hablar, vivir y hasta de pensar).

3.3.2. Ventajas y desventajas de la entrevista

La entrevista personal tiene algunas ventajas con relación a los otros procedimientos de aplicación de encuestas, entre las que se mencionan:

a) El tratamiento de temas complejos, dado que el entrevistador puede aclarar aspectos no entendidos.

b) Comprobar la comprensión de las preguntas y la consistencia de las respuestas.

c) Recoger información complementaria sobre las características del entrevistado, como su entorno familiar, la vivienda donde habita y otros aspectos relevantes.

d) Despertar el interés del entrevistado a través de técnicas de motivación.

e) La aclaración de dudas que surjan del entrevistado.

Entre las desventajas se pueden mencionar:

a) El costo de las entrevistas generalmente es más alto con relación a otras opciones.

b) La colaboración de la gente; hay personas que son reacias a colaborar.

c) El acceso a las viviendas por temas de inseguridad ciudadana, principalmente en grande áreas urbanas.

d) La localización de determinados grupos de población, sobre todo si el trabajo de campo se realiza en horarios que corresponden a la jornada laboral. De realizarse en estos horarios, se corre el riesgo de una sobrerepresentación de amas de casa, jubilados y desempleados (Cea D'Ancona 1999: 245).

3.4. Trabajo de campo

En el proceso de una encuesta, el trabajo de campo o de recogida de información es delicado y conlleva costos económicos. En esta actividad intervienen generalmente los encuestadores y los supervisores y de su labor dependerá la calidad de la información obtenida.

Los encuestadores tienen el papel más importante en el trabajo de campo; entre sus funciones se encuentran las siguientes:

a) Localizar y visitar a los entrevistados.

b) Motivar para conseguir la participación de los entrevistados.

c) Leer las preguntas en su exacta formulación y en el orden en que aparecen en el cuestionario.

d) Comprobar si las respuestas se adecuan a los propósitos de cada pregunta; en las preguntas abiertas se deberá registrar las respuestas lo más exactamente posible a la declaración del entrevistado.

Dadas las responsabilidades del entrevistador, es necesario su adiestramiento y capacitación en todos los temas relacionados con la encuesta, los que tienen que ver con:

a) El conocimiento de los objetivos de la encuesta.

b) Información sobre el marco muestral.

c) La revisión del cuestionario pregunta por pregunta, aclarando el significado de cada una de ellas, los códigos utilizados, la forma de llenado, el tratamiento a las no respuestas.

d) Conocimiento del diseño de campo sobre la ubicación de las unidades a entrevistar, los tiempos fijados para cada entrevista y los materiales que se deben portar.

e) Las formas de realizar la entrevista que tiene que ver con la presentación, aplicación del cuestionario, procedimientos para aumentar la tasa de colaboración y la despedida.

f) La relación con el supervisor para recibir orientación y revisar la calidad de los datos recogidos.

En encuestas de carácter nacional o de muestras muy grandes se sugiere preparar un manual para el encuestador en el que se especifican en forma detallada todas las actividades y responsabilidades del encuestador. Un buen ejemplo de manual es el preparado por el INE para las Encuestas de Hogares y Calidad de Vida del Programa MECOVI.

Para mejorar los resultados del entrevistador es necesario contar con una adecuada supervisión a fin de asegurar que los entrevistadores están aplicando correctamente la entrevista. Este monitoreo requiere que el supervisor realice las siguientes actividades:

a) Revise los cuestionarios llenados, comprobando que los datos de identificación, el porcentaje de respuesta, la omisión de preguntas, la codificación, se encuentran en los marcos establecidos previamente en el diseño.

b) Compruebe, cuando se ha establecido una ruta para la selección de las unidades, si el encuestador entrevista a quienes tiene que entrevistar; en caso de ausencias, ayudar al encuestador para seleccionar otra unidad que no afecte el diseño muestral.

c) Colaborar al encuestador en los distintos aspectos del trabajo de campo.

d) Elaborar un informe de supervisión en el que se refieran todos los aspectos relevantes del trabajo de campo y en el que recomiende, en caso de ser necesario, volver a entrevistar algunas unidades para corregir errores.

La planificación del trabajo de campo debe ser una tarea que involucre tanto las distintas fases de la entrevista para la aplicación del cuestionario como las responsabilidades de los encuestadores, de los supervisores y, cuando sea necesario, del jefe de campo. También debe incorporar la programación de los tiempos, el requerimiento de materiales y los costos de manera desagregada.

3.5. Codificación

Concluida la fase de recolección de la información, existe un gran número de datos individuales que deben ser ordenados y clasificados para su posterior análisis. Es decir, se debe pasar del dato bruto al dato procesado. La clasificación implica la codificación que consiste en asignar a todas y cada una de las categorías que comprende el cuestionario números o signos para hacer posible la agrupación de los datos. En muchos casos en los cuestionarios existe una precodificación en las preguntas cerradas (Sierra Bravo 2003: 408).

La codificación de las respuestas implica asignar un número a cada una de ellas, documentar el significado de los códigos, aplicar la codificación y guardar los datos en archivos magnéticos. Se asignan números sólo a

variables categóricas o cualitativas. Cada pregunta debe codificarse de manera correlativa con números u otros símbolos.

Ejemplo No. 25

Se puede codificar las preguntas y posibles respuestas de la siguiente manera:	
Si usted no trabaja, ¿cuál es la razón? (1) 1. = Estudiante 2. = Ama de casa 3. = Jubilado 4. = Incapacitado 5. = Rentista 6. = Otro (especifique) Se observa que el código de la pregunta es 1 y el de las repuestas se encuentra entre 1 y 6. La pregunta 1 puede tener valores entre 1 y 6.	*Desempeña su ocupación como: (5)* 1. = Empleado 2. = Obrero 3. = Patrón 4. = Cuenta propia 5. = Otro (especifique) Se observa que el código de la pregunta es 5 y el de las repuestas se encuentra entre 1 y 5. La pregunta 5 puede tener valores entre 1 y 5.

La asignación de códigos y sus procedimientos deben documentarse en un informe detallado que se constituya en una guía para el proceso de codificación y para localizar las variables e interpretar los datos durante el análisis. En este documento debe incorporarse la codificación de las no respuestas y su tratamiento en la clasificación (Hernández/Fernández/Baptista 1999: 320).

Efectuada la codificación que permite la clasificación de la información contenida en los cuestionarios, ésta se procede a guardarla en algún medio magnético.

4. Análisis de los datos

Una vez realizada la recolección de la información de tipo cuantitativo, comienza la fase del análisis e interpretación. Este proceso puede resultar complejo dependiendo del tipo de análisis que se pretenda realizar: descriptivo, inferencial, causal univariado o multivariado, pruebas paramétricas, entre los más importantes. El alcance de este apartado se orienta a presentar una introducción a los aspectos básicos sobre la presentación de la información y el análisis descriptivo.

4.1. Presentación de la información cuantitativa

Ya sea que se maneje información de una muestra o de una población, existe una regla práctica: cuando el conjunto de datos contenga veinte o más observaciones, la mejor manera de observar los datos es presentarlos de forma resumida elaborando tablas y gráficas apropiadas que faciliten aproximar sus principales características (Berenson/Levine 1992: 97).

4.1.1. Distribución de frecuencias

Es una tabla resumen en la que se ordenan los datos por clases o categorías. En cada clase o categoría existirá más de un dato denominado frecuencia que cumple con el atributo fijado en la clase o categoría. Debe tenerse en cuenta que al agrupar los datos se pierde alguna información de las observaciones individuales.

En la construcción de una tabla de frecuencias se debe considerar tres aspectos: La selección del número adecuado de clases o categorías, la definición del intervalo de clase, y los límites de cada clase a fin de evitar traslapes.

Selección del número de clases:

El número de clases o categorías para variables cuantitativas depende de la cantidad de datos; cuanto mayor es el número de observaciones, mayor el número de clases. Si existen pocas clases, la información que se obtiene es reducida y si hay muchas se tendrá una baja concentración de datos no permitiendo una buena descripción de éstos. Si las variables son cualitativas, el número de categorías está en función de los agrupamientos que previamente definió el investigador. Por ejemplo, para el caso de estado civil, las categorías pueden ser: soltero, casado, conviviente, separado, divorciado, viudo.

Intervalo de clase:

Es el tamaño o anchura de cada clase, sólo se aplica a variables cuantitativas. Para su cálculo, un camino es que todas las clases tengan el mismo tamaño, en este caso se busca la diferencia entre el dato mayor y el menor (denominado rango) y se divide entre el número de clases que se desean.

$$\textit{Tamaño de clase} = \frac{\textit{Rango}}{\textit{Número de clase}}$$

Si el resultado fuera un número decimal, se debe redondear al entero superior.

Una segunda opción se presenta cuando existen elevadas diferencias entre los datos y se desea analizar algunos grupos con especial atención; en esta situación los intervalos de clase no requieren ser del mismo tamaño.

Límite de clase:

Son las acotaciones para cada intervalo de clase que permiten agrupar los datos sin que éstos se sobrelapen, es decir que sólo pertenezcan a una clase. Existen intervalos cerrados, que son los que tienen un límite inferior y superior, y abiertos, en los que no se acota el límite superior o inferior ("más de" o "menos de").

Marca de clase:

Es el punto o valor que se encuentra en la mitad de los límites de cada clase y es representativo de los datos de esa clase. Se calcula de la siguiente manera:

$$\textit{Marca de clase} = \frac{\textit{Límite inferior} + \textit{Límite superior}}{2}$$

4.1.2. Distribución de frecuencia relativa y acumulada

A partir de la distribución de frecuencias, con el propósito de mejorar el análisis, se puede elaborar la distribución de frecuencias relativas que se obtiene dividiendo la frecuencia de cada una de las clases o categorías entre el número total de observaciones, el resultado será un número menor a uno y mayor a cero denominado proporción. La expresión de la frecuencia relativa se puede expresar en porcentajes multiplicando cada frecuencia por cien. La suma de las frecuencias relativas da lugar a las frecuencias acumuladas.

Ejemplo No. 26

Tomando como referencia el ejemplo donde se propone analizar los factores que inciden en el nivel de ingreso de los ocupados de la ciudad de La Paz, la Encuesta de Hogares-Calidad de Vida del programa MECOVI (INE) ha establecido una muestra para la ciudad de La Paz de 420 viviendas donde

(Continúa en la página siguiente)

(Viene de la página anterior)

viven 1.800 personas; de éstas, 506 son ocupadas y constituyen el objeto de estudio.

Distribución de frecuencias por grupos de edad:

Para elaborar esta tabla se observó en la muestra de ocupados que las edades están comprendidas entre los 10 y 80 años, estableciéndose un rango de 80 – 10 = 70. Se definió agrupar la población en cinco intervalos de clase (con propósito didáctico). Con esta información se calculó el intervalo o tamaño de clase.

$$\textit{Tamaño de clase} = \frac{\textit{Rango}}{\textit{Número de clase}} = \frac{70}{5} = 14$$

Los límites de clase se establecieron en base al tamaño de clase, el límite inferior de la primera clase fue de 10 años y el superior hasta menos de 24 años (23.999) (10 + 14). Con este mismo procedimiento se establecieron los límites para la segunda clase (24 hasta menos de 38 años) y así sucesivamente. Se evitó que existan traslapes entre los intervalos.

Definidos los aspectos anteriores se estructuró la siguiente tabla de frecuencias absolutas, relativas y acumuladas de la muestra de la población ocupada por grupos de edad.

Tabla de frecuencias por grupo de edad			
Grupo de edad	**Frecuencia absoluta**	**Frecuencia relativa (%)**	**Frecuencia relativa acumulada (%)**
10 – 23 años	113	22,3	22,3
24 – 37 años	170	33,6	55,9
38 – 51 años	134	26,5	82,4
52 – 65 años	69	13,6	96,0
66 – 80 años	20	4,0	100,0
Total	**506**	**100,0**	

Distribución de frecuencias por grupos de ingresos:

En la elaboración de esta tabla se observó que el rango tenía una amplitud de Bs. 35.675 y que existía una elevada dispersión de los datos, los que se

(Continúa en la página siguiente)

(Viene de la página anterior)

concentraban en ingresos inferiores a Bs. 2.000. Por esta razón, se optó en delimitar diferentes tamaños de clase que permitieran cubrir todo el universo y que el número de clases no fuera elevado; se optó por cinco grupos. Los resultados se muestran a continuación.

Tabla de frecuencias por grupo de ingreso			
Grupo de ingreso	**Frecuencia absoluta**	**Frecuencia relativa (%)**	**Frecuencia relativa acumulada (%)**
00 - 500	185	36,6	36,6
501 - 1.000	156	30,8	30,8
1.001 - 1.500	59	11,7	11,7
1.501 - 2.000	28	5,5	5,5
2.001 - 4.000	42	8,3	8,3
4.001 - 36.000	36	7,1	7,1
Total	**506**	**100,0**	**100,0**

4.1.3. Tablas de contingencia

Estas tablas relacionan o cruzan dos variables que pueden ser cuantitativas o cualitativas, lo cual permite, dependiendo del propósito de la investigación, tener una visión de los datos en forma más desagregada e inferir de manera preliminar un primer nivel de relación entre las variables.

4.1.4. Supertablas

Son tablas de contingencia donde se muestran tres o más variables, las que pueden ser una combinación de variables cuantitativas (previamente ordenas por clases o categorías) con variables cualitativas.

Ejemplo No. 27

Con base en los datos provenientes de la Encuesta de Hogares, se puede armar una tabla de contingencia que cruce los grupos de edad por sexo, los resultados se pueden mostrar en valores absolutos y relativos (porcentajes).

(Continúa en la página siguiente)

(Viene de la página anterior)

Tabla de contingencia Edad por Sexo									
Grupos de edad	Frec. Absolutas			Frec. Relativas Vertical			Frec. Relativas Horizontal		
	Sexo		Total	Sexo		Total (%)	Sexo		Total (%)
	Hombre	Mujer		Hombre (%)	Mujer (%)		Hombre (%)	Mujer (%)	
10 - 23	52	61	113	19,5	25,4	22,3	46,0	54,0	100,0
24 - 37	87	83	170	32,7	34,6	33,6	51,2	48,8	100,0
38 - 51	73	61	134	27,4	25,4	26,5	54,5	45,5	100,0
52 - 65	39	30	69	14,7	12,5	13,6	56,5	43,5	100,0
66 - 80	15	5	20	5,6	2,1	4,0	75,0	25,0	100,0
Total	**266**	**240**	**506**	**100,0**	**100,0**	**100,0**	**52,6**	**47,4**	**100,0**

En la tabla se puede observar (frecuencias relativas horizontal) que de la muestra de la población ocupada de la ciudad de La Paz, el 52,6% son hombres y el 47,4% son mujeres. Las frecuencias relativas verticales muestran, para el caso de los hombres, que las mayores participaciones (porcentajes) se presentan en los grupos de edad comprendidos entre los 24 a 37 años (32,7%) y entre los 38 a 51 años (27,4%), en el caso de las mujeres las mayores participaciones se presentan en los tres primeros grupos de edad. Las frecuencias relativas horizontales revelan que en el grupo de edad comprendido entre los 10 a 23 años la participación (porcentaje) de la mujer (54%) es mayor a la de los hombres (46%) mientras que en los dos últimos grupos de edad, la participación de los hombres es mayor.

Una tabla de contingencia más amplia (Supertabla) podrá ser aquella que cruce los grupos de ingreso con grupos de edad clasificados por sexo.

Tabla de contingencia Grupos de Ingreso por Grupos de Edad por Sexo											
Grupos de ingreso	Grupos de edad por sexo										Totales
	10 - 23		24 - 37		38 - 51		52 - 65		66 - 80		
	Hombre	Mujer	Hombre	Mujer	Hombre	Mujer	Hombre	Mujer	Hombre	Mujer	
00 - 500	27	36	19	42	9	22	11	12	5	2	185
501 - 1.000	18	19	33	21	21	19	9	11	4	1	156
1.001 - 1.500	5	4	16	5	11	9	6	1	1	1	59
1.501 - 2.000	2	2	6	3	7	3	2	2	1		28
2.001 - 4.000			5	9	13	1	7	3	3	1	42
4.001 - 36.000			8	3	12	7	4	1	1		36
Totales	**52**	**61**	**87**	**83**	**73**	**61**	**39**	**30**	**15**	**5**	**506**

4.2. Representaciones gráficas

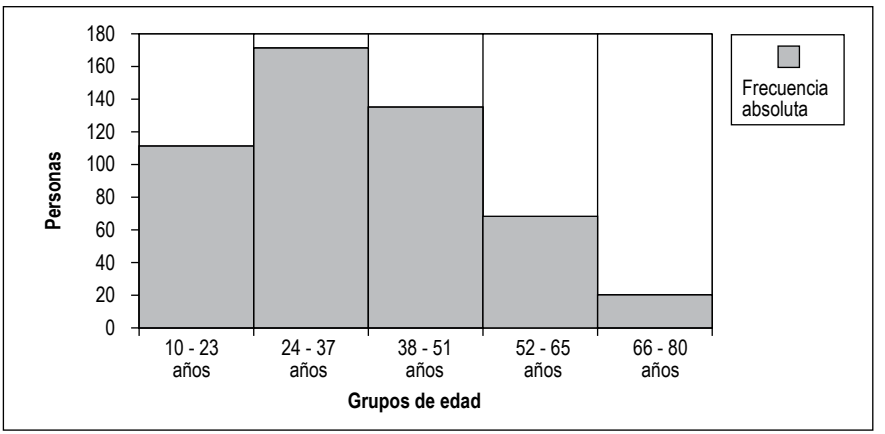
La información ordenada en tablas de frecuencia o de contingencia o cuando se trata de un número reducido de datos se puede representar en forma de gráficas, lo cual permite destacar aspectos significativos de un conjunto de datos, bien se dice "una imagen vale más que mil palabras". Las formas más comunes de representar los datos son los histogramas, los polígonos, diagramas de barra, de pastel, radiales.

4.2.1. Los histogramas

Son una forma de representación gráfica en dos ejes mediante barras verticales, cuyo ancho está definido por los límites de clase. En el eje horizontal se representa las categorías (intervalos de clase), por ejemplo, los intervalos de edad o de ingreso, y en el eje vertical se representan el número, proporción o porcentaje de observaciones para cada intervalo de clase. La altura de cada barra representa la frecuencia de cada clase o categoría. Si en el eje vertical se representa los valores absolutos, el histograma se denomina de frecuencias absolutas; si son porcentajes, se denomina de frecuencias relativas o de proporciones.

De la tabla de frecuencias del Ejemplo No. 26 de la población ocupada por grupos de edad se puede graficar un histograma de frecuencias absolutas que tendrá la siguiente forma:

Histograma
Frecuencias Absolutas

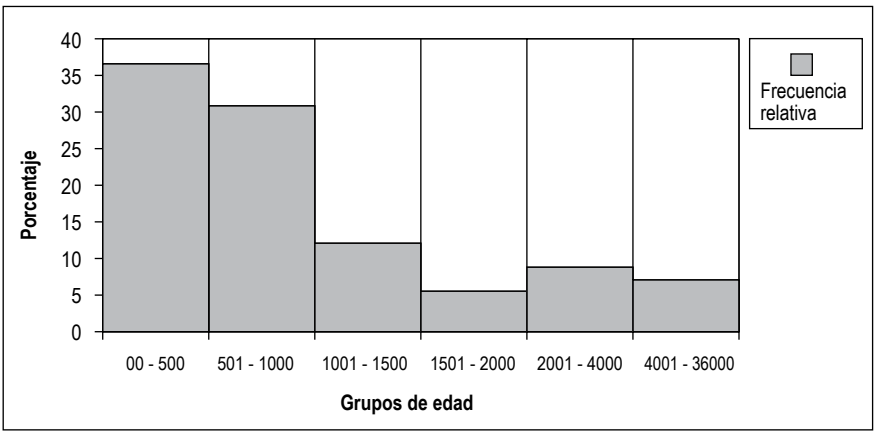


En el histograma se aprecia una aproximación de la distribución de la población ocupada por grupos de edad. Se observa que el mayor número de personas ocupadas se encuentra en el grupo de edad comprendido entre 24 a 37 años (170 personas), le sigue el grupo de 38 a 51 años (134 personas) y el de 10 a 23 años (113 personas), mientras que en los grupos comprendidos entre 52 y 65 años y 66 a 80 años se encuentran en cada uno menos de 70 personas.

Esta distribución permite inferir que más de 80% de los ocupados se encuentran en los tres primeros grupos de edad (entre 10 y 51 años), que al aumentar la edad hasta los 37 años crece la ocupación para luego disminuir paulatinamente conforme la población se hace más vieja. El mayor número de ocupados entre el segundo y tercer grupo puede responder a varios factores, como un mayor grado de educación y experiencia, la necesidad de generar ingresos para la familia dado que la mayor parte de esta población no son solteros.

También se puede graficar un histograma de frecuencias relativas por grupos de ingreso contenida en el Ejemplo No. 26:

Histograma
Frecuencias Relativas



En este histograma se observa, en base a la muestra considerada en el ejemplo, que los ingresos de la mayor parte de la población ocupada (68%) no sobrepasan Bs. 1.000, en efecto 37% de la población ocupada tiene un ingreso hasta de Bs. 500 y 31% entre Bs. 501 y Bs. 1.000. En los siguientes grupos, que corresponden a ocupados con cada vez mayores de niveles de ingreso, su representatividad en el total de la muestra se

reduce conforme se incrementa el ingreso. Así se observa que en el grupo de mayores ingresos se encuentra 7% de los ocupados. Este histograma permite una primera aproximación a la distribución del ingreso entre los ocupados, destacándose que la mayor parte de los ocupados tienen bajos ingresos y sólo 15% tiene ingresos por encima de Bs. 2.000.

4.2.2. Polígonos de frecuencia

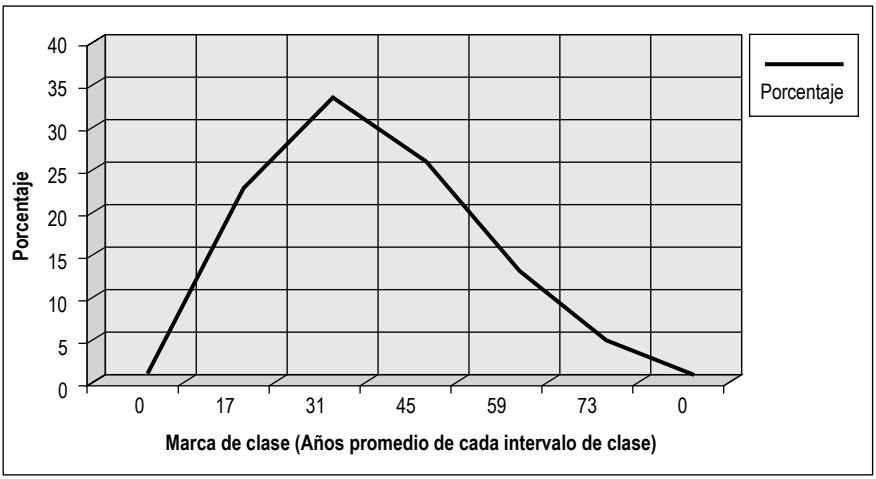
Son otra forma de representación gráfica de línea continua que se construye de modo análogo al histograma, representado en el eje horizontal las marcas de clase y en el vertical las frecuencias absolutas o relativas. Cuando se trata de propósitos comparativos entre dos poblaciones o muestras, se recomienda utilizar las frecuencias relativas.

En la construcción de los polígonos se debe tener presente que es la representación de la forma de una distribución en particular. El área debajo del polígono incluye la totalidad de las observaciones sea que se expresen en valores absolutos o relativos, por lo que es necesario conectar los puntos medios primero y último con el eje horizontal. Asimismo, el eje vertical debe mostrar el cero para no distorsionar o representar en forma equivocada los datos.

Los polígonos permiten una primera aproximación a la forma de distribución de los datos. Cuando su forma es achatada implica la existencia de dispersión de los datos, en tanto que cuando su forma es menos achatada y apuntalada, los datos se concentran alrededor de una medida de tendencia central. Asimismo, se puede ilustrar las propiedades de tendencia central, dispersión y forma.

De la tabla de frecuencias por grupos de edad sobre la población ocupada se puede graficar un polígono de frecuencias relativas que tendrá la siguiente forma:

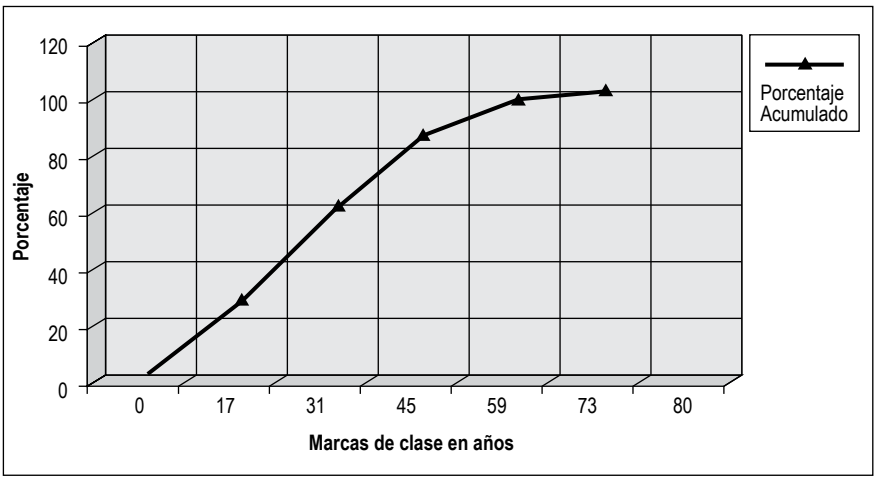
Polígono
Frecuencias Relativas



En este polígono se aprecia que más de 30% de la población ocupada tiene una edad de alrededor de 31 años y que no existe una elevada dispersión de las edades respecto a este valor. Asimismo, se observa que la distribución porcentual de las edades se aproxima a una curva normal con cierto sesgo hacia la derecha.

También se puede graficar un polígono de frecuencias acumuladas denominado ojiva. Considerando el ejemplo anterior, en el eje horizontal se representan las marcas de clase y en el vertical los porcentajes acumulados. Esta información proviene de la tabla de frecuencias acumuladas del Ejemplo No. 26.

Polígono
Frecuencias Acumuladas (Ojiva)



En la gráfica se observa que más de 95% de la muestra de ocupados de la ciudad de La Paz tiene hasta 59 años, que los ocupados con menos de 40 años representan más de 60% y los de menos de 17 años un poco más de 20%.

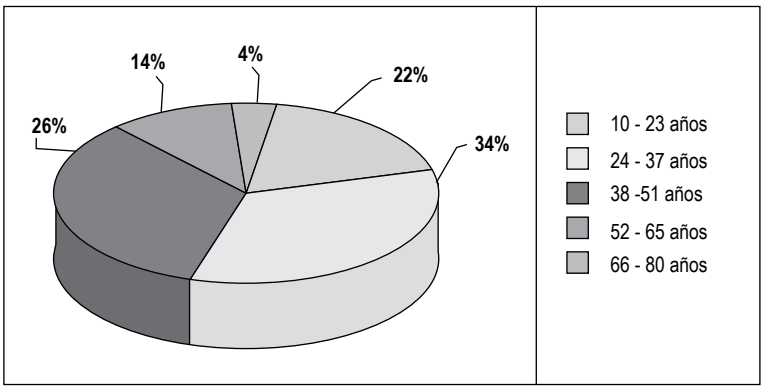
4.2.3. Gráficas de pastel

Este tipo de gráficas son utilizadas para mostrar proporciones o porcentajes de participación de una categoría o intervalo de clase en la población o muestra total que es objeto de estudio. Se expresa a través de un círculo que se divide en proporciones que corresponden al porcentaje con que cada categoría participa en el universo seleccionado (población o muestra).

La gráfica de pastel, denominada también "torta", permite visualmente detectar la representatividad de cada categoría en el universo sujeto de estudio.

Mediante una gráfica de pastel o torta y en base a la tabla de frecuencias relativas del Ejemplo No. 26 se puede representar la proporción (porcentaje) con que cada grupo de edad de la población ocupada de la muestra de la ciudad de La Paz participa en la misma.

Gráfica de pastel o torta
(% de participación de cada grupo de edad en la población ocupada)



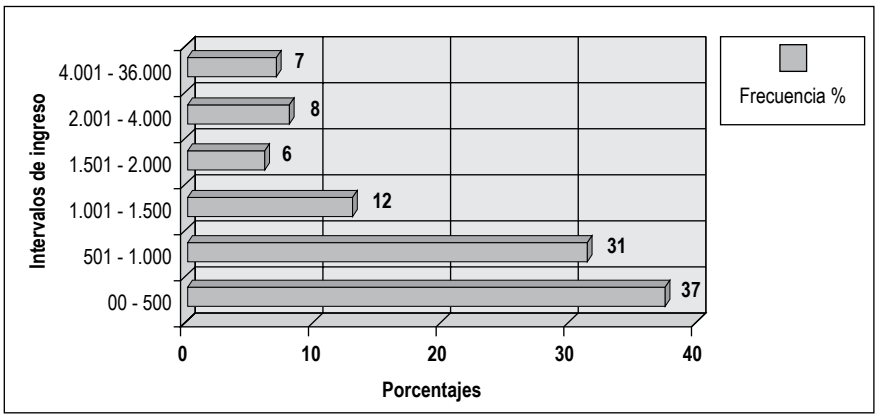
La gráfica permite una aproximación al perfil etáreo de la población ocupada, destacando que los mayores porcentajes corresponden a los tres primeros grupos de edad, comprendidos entre los 10 a 51 años (22%, 34% y 26%, respectivamente), en tanto que los grupos con más de 51 años muestran porcentajes de participación reducidos. Llama la atención que en el grupo de 66 a 80 años (edad de jubilación) se encuentren personas ocupadas, situación que puede deberse a la falta de rentas o de redes de seguridad familiar.

4.2.4. Gráfica de barras

Una forma alternativa de mostrar la representatividad de una categoría en la población total o muestra es la gráfica de barras, que puede ser de barras horizontales o verticales. Esta gráfica se construye en dos ejes, cada barra representa una categoría (intervalo de clase) que debe tener el mismo ancho para no confundir al lector. Si se opta por la forma horizontal, en el eje vertical se colocan las categorías y en el horizontal las frecuencias absolutas o relativas (porcentajes); si se prefiere la forma vertical, se debe invertir el significado de los ejes.

Para mostrar las frecuencias relativas de la muestra de la población ocupada por intervalos de ingreso la gráfica de barras es la siguiente:

Gráfica de barras
(% de ocupación en cada grupo de ingreso)



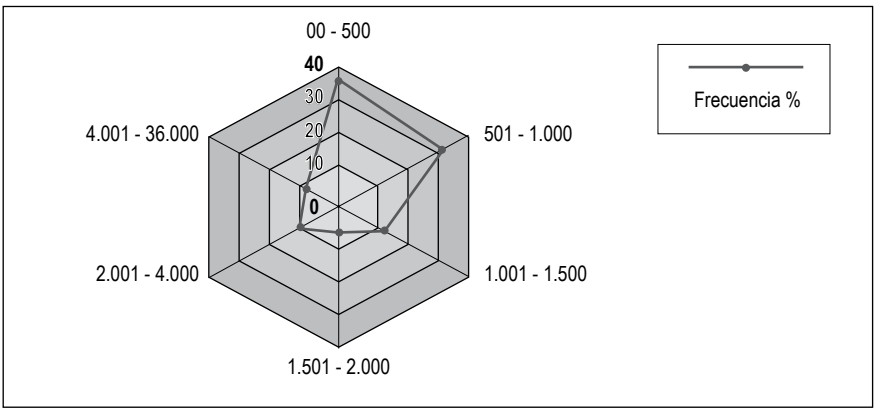
Esta gráfica ratifica los comentarios efectuados en el histograma de distribución de frecuencias relativas: elevada concentración de la población ocupada con ingresos menores a Bs. 1.001 y reducidos porcentajes de participación de ocupados con ingresos elevados. Si se considera que los ingresos de los ocupados se distribuyen al interior de la familia y dada la situación que nos muestra la gráfica se puede explicar en parte la pobreza que se verifica en la ciudad de La Paz.

4.2.5. Gráfica radial

Este tipo de gráficos se utilizan generalmente para comparar dos distribuciones de frecuencias relativas cuando se utiliza ya sea dos poblaciones o muestras o se quiere representar los cambios de una distribución de frecuencias en dos a más momentos en el tiempo. En esta gráfica los ejes son polígonos concéntricos, donde cada polígono representa las frecuencias relativas, y los vértices externos muestran las categorías o intervalos de clase.

La distribución de frecuencias relativas de la población ocupada por intervalo de ingreso se representa en una gráfica radial de la siguiente forma:

Gráfica radial
(Participación (%) de la población ocupada en cada grupo de ingreso)



Todas las representaciones gráficas contenidas en este texto y otras no incluidas, pueden elaborarse utilizando paquetes computacionales que contienen una amplia gama de posibilidades para elaborar gráficos. Dado el amplio uso de hojas de trabajo como Excel o Lotus, éstas incorporan variadas opciones para graficar a partir de ordenamientos de datos. Los paquetes estadísticos como el SPSS o STATA también tienen una amplia biblioteca de gráficas.

4.3. Análisis descriptivo de datos

Tanto la construcción de tablas de frecuencias, de contingencia y supertablas, así como la presentación gráfica de los datos permite una primera visión general de algunos aspectos destacables de la información; sin embargo, es necesario pasar a otro nivel que permita mejorar la descripción del conjunto de datos cuantitativos que son objeto de una investigación.

La descripción de los datos se puede realizar en consideración a su tendencia central, dispersión y forma, para lo cual se utilizan distintas medidas de carácter descriptivo. Si estas medidas se calculan a partir de una muestra de datos se denominan "estadísticos", y si se calculan a partir de una población se denominan "parámetros".

4.3.1. Medidas de tendencia central

La mayor parte de los conjuntos de datos muestran una tendencia a agruparse alrededor de un punto central y, por lo general, es posible

elegir algún valor promedio que describa todo un conjunto de datos. Un valor típico descriptivo como ése es una medida de tendencia central o posición. Las medidas más utilizadas son: la media aritmética, mediana y la moda.

La media aritmética:

Es el promedio o medida de tendencia central que se calcula sumando todas las observaciones de un conjunto de datos, dividiendo después ese total entre el número total de elementos de la población o muestra. La fórmula que se utiliza para el cálculo de la media aritmética para una muestra es la siguiente:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Donde; $\sum_{i=n}^{n} X_i = X_1 + X_2 + \ldots\ldots + X_n$

El símbolo $\sum$ se lee como sumatoria de todos los valores
n = Número de observaciones en la muestra
X_i = i–ésima observación de la variable X.

Esta fórmula es aplicable cuando se cuenta con pocos datos o se dispone de una computadora. Cuando el número de datos es grande, se los agrupa en tablas de frecuencia y el promedio se denomina "media aritmética ponderada", cuya fórmula es:

$$\overline{X} = \frac{\sum_{j=1}^{g} m_j f_j}{n}$$

Donde: g = número de clases o grupos
m_j = punto medio o marca de clase de la j-ésima clase
f_j = número de observaciones en la j-ésima clase o frecuencia de la j-ésima clase

El cálculo de la media se basa en todas las observaciones del conjunto de datos; ninguna otra medida de tendencia central tiene esta característica. Sin embargo, como su cálculo se basa en todas las observaciones, resulta muy afectada por valores extremos. Cuando se presenta esta situación, la media aritmética puede representar una imagen distorsionada y no es el mejor indicador para describir un conjunto de datos.

Debe aclararse que el cálculo de la media a partir de datos no agrupados (datos en su forma original) ofrece resultados reales y su cálculo con datos agrupados (tablas de frecuencias) son resultados muy aproximados a la media real.

Ejemplo No. 28

De un grupo de seis personas de 18, 22, 25, 28, 31, 33 años, se quiere conocer el promedio de estas edades.

La media será:

$$\bar{X} = \frac{18 + 22 + 25 + 28 + 31 + 33}{6} = 26{,}16 \text{ años} \approx 26 \text{ años}$$

Si los datos se encuentran en una tabla de frecuencia (como la del Ejemplo No. 26 sobre las edades de la población ocupada), el cálculo de la media implica:

- Determinar el punto medio (marca de clase) de cada intervalo. En este cálculo se incluye el límite inferior y el límite superior considerando que éste es menor al límite inferior de la siguiente clase (10 + 23.999)/2 = 16.999, valor que puede redondearse a 17.
- Cada marca de clase se multiplica por su frecuencia absoluta.
- La suma de esta multiplicación se divide entre la suma de las frecuencias absolutas que es el tamaño de la muestra o de la población. Este resultado es la media ponderada.

Grupo de edad	Marca de clase	Frecuencia absoluta	m * f
10 - 23 años	17	113	1.921
24 - 37 años	31	170	5.270
38 - 51 años	45	134	6.030
52 - 65 años	59	69	4.071
66 - 80 años	73	20	1.460
Total		**506**	**18.752**

$$= \frac{\sum_{j=1}^{g} m_j f_j}{n}$$

$\bar{X} = \frac{18.752}{506} = 37 \text{ años}$, que es la medida ponderada de la muestra de la población ocupada de la ciudad de La Paz.

(Continúa en la página siguiente)

(Viene de la página anterior)

El cálculo de la media ponderada de los ingresos de la muestra de los ocupados de la ciudad de La Paz, siguiendo el mismo procedimiento, para el año 2001 fue de Bs. 2.237 al mes. Este promedio está fuertemente influenciado por pocos casos (6) de ingresos que sobrepasan los Bs. 12.000. Sin considerar estos casos la media fue de Bs. 1.383.

La mediana:

Es el valor que se encuentra en el centro de un conjunto ordenado de datos y divide en dos el conjunto. Es decir que la mitad de los datos están por debajo de la mediana y la otra mitad por encima. La mediana no se ve afectada por observaciones extremas, por ello, cuando existe alguna observación extrema, resulta apropiado utilizar la mediana.

Para el cálculo de la mediana lo primero que se debe hacer es ordenar los datos en forma ascendente o descendente. Si el número de datos es impar, se aplica la siguiente fórmula:

$$Mediana = \frac{n + 1}{2}$$

Si el número de datos es par también se aplica la fórmula anterior, pero entonces la posición de la mediana estará entre dos observaciones intermedias, en tal caso se calculará el promedio correspondiente a estas dos observaciones centrales. Si en la muestra se repiten valores, en el cálculo de la mediana se ignora esta situación y se calcula en base a las referencias anteriores.

El cálculo de la mediana se afecta por la cantidad de observaciones y no por la magnitud de ningún valor extremo.

Ejemplo No. 29

En base al Ejemplo No. 28 referido a la edad de 6 personas, se tiene el siguiente arreglo:

Persona	1	2	3	4	5	6
Edad	18	22	25	28	31	33

$$Mediana = \frac{n+1}{2} = \frac{6+1}{2} = 3{,}5$$ valor que se encuentra entre 25 y 28 años; aplicando el promedio (25 + 28)/2 = 26,5 años se encuentra la mediana, que representa que el 50% de las personas son menores a esta edad y el otro 50% son mayores. Debe observarse en este caso que los valores de la media (26 años) y la mediana son muy próximos.

La mediana de ingresos de la muestra de los ocupados de la ciudad de La Paz es de Bs. 703; en este caso, se pudo apreciar una importante diferencia con la media aritmética, diferencia que se debe a la presencia de valores extremos elevados. Por esta característica se recomienda utilizar para este caso la mediana.

La mediana para datos agrupados:

Para datos agrupados en frecuencias de clase, el cálculo de la mediana se realiza con la siguiente fórmula:

$$Mediana = L + \left(\frac{n/2 - F}{f_{med}} \right) c$$

Donde:
- L = Es el límite inferior del intervalo de clase donde se encuentra la mediana
- n = Tamaño de la muestra
- F = Es la suma de las frecuencias acumuladas hasta la clase de la mediana pero sin incluirla
- f_{med} = Es la frecuencia de la clase donde se encuentra la mediana
- c = Es el ancho del intervalo de clase donde se encuentra la mediana.

Ejemplo No. 30

Con base en la tabla de distribución de frecuencias del ingreso de los ocupados (Ejemplo No. 26),

Grupos de ingreso	Frecuencia absoluta	Frecuencia absoluta acumulada
00 - 500	185	185
501 - 1.000	156	341
1.001 - 1.500	59	400
1.501 - 2.000	28	428
2.001 - 4.000	42	470
4.001 - 36.000	36	506
Total	**506**	

El cálculo de la mediana será:

$$Mediana = L + \left(\frac{n/2 - F}{f_{med}}\right) c = 501 + \left(\frac{\frac{561}{2} - 185}{341}\right) 500 = 599{,}7 \cong 600\ Bs$$

En el procedimiento primero se debe calcular n/2 (506/2 = 253). Este valor permite ubicar en la columna de frecuencias acumuladas el grupo de ingreso donde se encuentra la mediana. Dado que 253 es mayor a 185 y menor a 341, la categoría de ingreso corresponde al intervalo comprendido entre 501-1.000. Luego, siguiendo la fórmula, se calculan los valores para F, c y f_{med}.

Debido a que los intervalos de ingreso no son del mismo tamaño, la mediana estimada a partir de datos agrupados difiere en más de 10% de la mediana calculada con los datos sin agrupar. Por esta razón, cuando los intervalos de clase no tienen el mismo ancho, se recomienda calcular la mediana a partir de los datos originales. Los paquetes estadísticos permiten esta opción en un tiempo muy reducido.

El cálculo de la mediana de edades de la población ocupada en la ciudad de La Paz, con el anterior procedimiento, es de 35 años y debido a que los intervalos de clase son iguales este resultado está más próximo a la mediana calculada con datos sin agrupar.

La moda:

Es el valor de un conjunto de datos que aparece con mayor frecuencia y se la obtiene a partir de un conjunto ordenado de datos. Para datos agrupados, la clase modal es el intervalo de clase con la frecuencia más alta y el valor modal por lo general se lo aproxima como el punto medio de esa clase modal. Esta medida no está afectada por valores extremos. Un conjunto de datos puede tener más de una moda o no tener moda cuando cada uno de los datos tiene la misma frecuencia.

Ejemplo No. 31

Considerando la tabla de distribución de frecuencias de la población ocupada por grupos de edad, el intervalo de clase que tiene el mayor número de frecuencias es el comprendido entre 24 y 37,99 años, que constituye la clase modal, y su valor modal es 31 años.

En la tabla de frecuencias de la muestra de la población ocupada por grupo de ingresos la moda se encuentra en el intervalo comprendido entre 0 y 500 Bs.

4.3.2. Medidas de dispersión

En la mayoría del conjunto de datos no todos los valores son iguales; el grado en que varían es de suma importancia para la investigación y la estadística en particular. Si no hubiera variación o dispersión de datos, no habría necesidad de la mayoría de las medidas de tendencia central y de otras que se utilizan en la estadística descriptiva. Las medidas de dispersión se orientan a describir la variabilidad de un conjunto de datos (Freund/Simon 1994: 70). Las más importantes son el rango, la varianza y la desviación estándar.

El rango:

Es la diferencia entre el valor más alto y el más bajo de un conjunto de datos. Se calcula mediante la siguiente fórmula:

$$Rango = X_{mayor} - X_{menor}$$

El rango mide la dispersión total del conjunto de datos. Su debilidad es que no toma en consideración la forma en que se distribuyen los datos entre los valores más pequeños y los más grandes. El cálculo del rango permite una primera aproximación a la dispersión de datos.

La varianza:

Es una medida de dispersión respecto a la media; mide qué tan cerca o tan lejos están los valores de su propia media aritmética y posibilita establecer la forma en que los valores fluctúan respecto al promedio.

Para calcular la varianza de una muestra de datos no agrupados se utiliza la siguiente fórmula:

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n - 1}$$

Donde: $\overline{X}$ = media aritmética de la muestra

n = tamaño de la muestra

X_i = i-ésimo valor de la variable X

La varianza para datos agrupados (en tablas de frecuencia) se calcula como:

$$S^2 = \frac{\sum_{j=1}^{g} (m_j - \overline{X})^2 f_i}{n - 1}$$

Donde: g = número de clases o grupos

m_j = punto medio o marca de clase de la j-ésima clase

f_i = número de observaciones en la j-ésima clase o frecuencia de la j-ésima clase

Los resultados que se obtienen con la varianza son unidades al cuadrado, por ejemplo, la edad al cuadrado o ingresos al cuadrado; a fin de expresar en unidades simples se utiliza la desviación estándar.

Desviación estándar:

Es una medida de dispersión respecto a la media aritmética que se obtiene aplicando la raíz cuadrada a la varianza.

Para datos no agrupados se calcula de la siguiente manera:

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n - 1}}$$

Para datos agrupados se emplea la siguiente fórmula:

$$S = \sqrt{\frac{\sum_{j=1}^{g} (m_j - \overline{X})^2 f_j}{n - 1}}$$

Los resultados que se obtienen para la varianza y la desviación típica permiten detectar que cuanto mayores son sus valores los datos están más dispersos respecto a su media. Si estos valores son pequeños, los datos estarán más concentrados alrededor de la media; si la varianza y la desviación típica son iguales a cero, significa que todas las observaciones son iguales.

Esta característica ayuda a verificar la homogeneidad o heterogeneidad de los grupos. Casi siempre, cuando se analiza aspectos relacionados con el ingreso, se observa elevada dispersión, producto de las diferencias de ingreso entre una población.

En el marco del teorema de Chebyshev, cuando se conoce la desviación estándar se puede establecer que 75% de la población se encuentra comprendida entre más dos y menos dos desviaciones estándar respecto a la media ($2\sigma + \mu - 2\sigma$), y 88,9% entre más tres y menos tres desviaciones estándar respecto a la media.

Ejemplo No. 32

Con base en la tabla de frecuencias de los grupos de edad de la población ocupada, se puede obtener la varianza y desviación estándar para datos agrupados. Para este cálculo se busca la marca de clase, luego las diferencias entre la marca de clase y la media; estas diferencias se elevan al cuadrado y este resultado se multiplica por las frecuencias. Esta secuencia se muestra en el siguiente cuadro:

De acuerdo al Ejemplo No. 28, la media de edad ($\overline{X}$) es de 37 años.

Grupo de edad	Marca de clase	$m_j - \overline{X}$	$(m_j - \overline{X})^2$	Frecuencia f_i	$(m_j - \overline{X})^2 f_i$
10 - 23 años	17	17-37= -20	400	113	45.200
24 - 37 años	31	31-37= - 6	36	170	6.120
38 - 51 años	45	45-37= 8	64	134	8.576
52 - 65 años	59	59-37= 22	484	69	33.396
66 - 80 años	73	73-37= 36	1296	20	25.920
Total				**506**	**119.212**

(Continúa en la página siguiente)

(Viene de la página anterior)

$\sum_{j=l}^{5} (m_j - \overline{X})^2 f_i = 11.9212$, luego la varianza será $S^2 = \frac{11.9212}{506\text{-}1} = 236$

y la desviación estándar $S = \sqrt{236} = 15{,}36$ *años*

En el ejemplo del ingreso de los ocupados, la desviación estándar es de Bs. 2.668.

Coeficiente de variación:

Es una medida relativa de dispersión que mide el porcentaje de variación respecto a la media. Es útil cuando se compara la variabilidad de dos o más conjuntos expresados en diferentes unidades de medición. La fórmula de cálculo es:

$$CV = \left(\frac{S}{\overline{X}} \right) 100\%$$

Donde: S = Es la desviación estándar de un conjunto de datos
$\overline{X}$ = Es la media de un conjunto de datos

Ejemplo No. 33

El coeficiente de variación para el caso de las edades de la población ocupada será

$$CV = \left(\frac{S}{\overline{X}} \right) 100\% = \left(\frac{15{,}36}{37} \right) = 41{,}51\%$$

El coeficiente de variación para los ingresos de los ocupados será

$$CV = \left(\frac{S}{\overline{X}} \right) 100\% = \left(\frac{2.668}{1.350} \right) = 100 = 197{,}6\%$$

(la media y la desviación estándar provienen de datos no agrupados)

Comparando estos resultados se aprecia que la dispersión de la información respecto a la media de las edades es mucho menor a la que se observa con el ingreso; en este caso, esta medida permite inferir la existencia de fuertes diferencias del ingreso respecto a la media.

4.3.3. La forma

Es la manera en que se distribuyen los datos. Una distribución de datos puede ser simétrica cuando tiene la forma de una campana cuyo centro divide en dos partes iguales los datos que se encuentran a su derecha e izquierda. Existe la distribución normal y la asimétrica o sesgada cuando no tiene la característica anterior.

Para describir la forma se compara la media y la mediana. Si estas dos medidas son iguales, se considera que los datos son simétricos (o con sesgo cero). Si la media es superior a la mediana, los datos tienen un sesgo positivo a la derecha. Si la media es menor a la mediana, los datos tienen un sesgo negativo o hacia la izquierda, luego:

Media >mediana =>sesgo positivo o hacia la derecha
Media <mediana =>sesgo negativo o hacia la izquierda
Media = mediana =>simetría o sesgo cero

Una medida de asimetría es el coeficiente de Pearson, que se calcula con la siguiente fórmula:

$$SK = \frac{3\ (media - mediana)}{\text{desviación estándar}}$$

Si el valor es negativo, existe sesgo hacia la izquierda; si es positivo, el sesgo es hacia la derecha; y si es cero, la distribución es simétrica.

El sesgo positivo se presenta cuando la media se ve afectada por valores muy grandes, por ejemplo, cuando se presentan ingresos muy elevados. El sesgo negativo ocurre cuando la media se reduce por valores muy pequeños, es el caso de la edad cuando predomina una población de niños y jóvenes. Los datos son simétricos cuando no hay valores extremos en dirección alguna, de manera que los valores bajos y altos se compensan entre sí.

4.3.4. Resumen

Tanto las medidas de tendencia central (la media, mediana y moda) como las de dispersión (rango, varianza, desviación estándar) y las de simetría permiten resumir la naturaleza general de un conjunto de datos. Con sólo pocas medidas se puede tener una idea razonablemente clara de una muestra o población.

De las medidas de tendencia central se obtiene una visión sobre dónde se localiza el conjunto de datos y si es simétrico o sesgado. De las medidas de dispersión se obtiene una idea de la variación de los datos respecto a la media, que también puede interpretase como la homogeneidad o heterogeneidad que puede presentar un determinado grupo respecto a algunas categorías que se desean investigar.

Ejemplo No. 34

A partir de las medidas descriptivas, se puede tener una visión sobre la posible relación entre el nivel de ingresos de los ocupados de la ciudad de La Paz, con la edad (proxy de la experiencia) y nivel de escolaridad.

Grupo de ingreso	Frecuencia	Promedio de edad	Coeficiente de variación edad %	Años de Escolaridad promedio	Coeficiente variación Escolaridad %
00 - 500	185	33,6	48,9	8,4	53,4
501 - 1.000	156	36,5	38,3	8,9	53,9
1.001 - 1.500	59	39,1	33,9	10,8	40,8
1.501 - 2.000	28	40,6	31,8	11,5	30,7
2.001 - 4.000	42	46,5	30,0	14,1	27,8
4.001 - 36.000	36	42,9	22,7	16,3	8,6
Total	**506**				

En el cuadro se observa que conforme aumenta el ingreso, el promedio de edad de las personas ocupadas es cada vez mayor, sin embargo, se observa la excepción en el grupo de los más altos ingresos. Esta situación invita a revisar las características de este grupo, comparar con una muestra similar y analizar caso por caso en la muestra a fin de validar o no este resultado.

Respecto a los coeficientes de variación para la edad, se aprecia que en el grupo de más bajos ingresos la dispersión es la más elevada, lo cual permite detectar que su composición etárea es heterogénea. En el grupo de los más altos ingresos se observa una menor dispersión respecto a su media de edad.

También se destaca una clara relación entre el nivel de ingreso y la escolaridad, mientras que en el grupo de ingresos bajos ésta es en promedio de 8,4 años, en el de más altos es de 16,3 años. Los coeficientes de variación de la escolaridad disminuyen a medida que aumenta el ingreso, que puede explicarse por el elevado número de ocupados por cuenta propia con distintos niveles de escolaridad que se encuentran en los tres primeros grupos, mientras que en los grupos de ingreso más elevado los aspectos de la formalidad en el trabajo exigen requisitos de escolaridad.

4.3.5. Cuantiles

Además de las medidas de tendencia central, dispersión y forma, existen otras medidas que se utilizan para resumir o descubrir propiedades de grandes conjuntos de datos cuantitativos. Los cuantiles dividen la población en proporciones iguales; así, los deciles dividen la población en diez partes, los cuartiles en cuatro, los quintiles en cinco. En este último caso, cada parte representa 20% de la población. El número de cuantiles dependerá del detalle que se quiera conocer de la población o muestra.

La construcción de los cuantiles implica los siguientes pasos:

a) Ordenar los datos de manera ascendente o descendente en función de la variable de interés, por ejemplo, para el caso de análisis de los ingresos se ordena a la población de los ingresos más bajos a los más altos.

b) Definir el número de cuantiles que se utilizarán para el análisis.

c) Crear los cuantiles de la población dividiendo el arreglo ordenado en las partes iguales que se han definido. Por ejemplo, si son quintiles de ingreso, se divide la población ordenada en cinco partes iguales, tal que cada una represente 20% de la población.

d) Luego, para cada cuantil se pueden aplicar medidas de tendencia central, de dispersión o forma.

Ejemplo No. 35

Se ha ordenado la muestra de los ocupados de la ciudad de La Paz por ingresos de manera ascendente y se ha definido agruparla por quintiles. Realizada esta agrupación, se han calculado algunas medidas descriptivas para el ingreso, la edad y la escolaridad.

Quintiles de ingreso	Ingreso promedio mensual Bs	Promedio de edad	Años de escolaridad promedio
1º	55	30,7	8,3
2º	394	36,9	8,1
3º	703	36,6	8,9
4º	1.116	38,1	10,3
5º	4.492	43,9	14,4

(Continúa en la página siguiente)

(Viene de la página anterior)

Esta presentación es muy común para análisis de ingresos y de pobreza y constituye una alternativa al cuadro de frecuencias por intervalos de ingreso que se ha mostrado en los ejemplos anteriores.

En el cuadro se destacan tendencias entre los ingresos, la edad y la escolaridad. Así, a medida que nos ubicamos en quintiles más elevados de ingreso, tanto la edad como la escolaridad van en aumento. Además, se observa que la diferencia entre el promedio de edad del quintil más bajo y el más alto es de 13 años y la diferencia de escolaridad entre estos quintiles es de 6,1 años.

Estos resultados exploratorios permiten adelantar que la hipótesis formulada sobre "Los mayores niveles de ingresos de los ocupados se explican por su nivel de escolaridad y por su experiencia" se verifica de manera preliminar.

Los siguientes pasos para ratificar o no estos resultados, obtenidos a partir de una muestra, requieren el proceso de verificar su validez externa (inferencia estadística) y aplicar las denominadas pruebas de hipótesis.

4.3.6. Inferencia estadística

El alcance del presente texto no aborda los temas de la inferencia estadística, que implica un conocimiento básico de la teoría de probabilidades, las distribuciones aleatorias y las pruebas de hipótesis, entre otros aspectos.

Sin embargo, debe tenerse presente la importancia de la inferencia estadística, que es el proceso de hacer uso de los resultados de una muestra para obtener conclusiones sobre las características de una población. Este proceso ayuda a determinar la representatividad de una muestra respecto a una población dada y los márgenes de error que se puedan tener al estimar los parámetros poblacionales.

Bibliografía

Abad de Servin, A. y A. Servin Andrade
1981 *Introducción al muestreo*. México: Limusa.

Alvira, F.
1989 "Diseños de investigación social: Criterios operativos", *Análisis de la realidad social*. F. Alvira, J. Ibañez y M. García Ferrando, comp. Madrid: Alianza.

Ander-Egg, E.
1972 *Introducción a las técnicas de investigación social*. Buenos Aires: Humanitas.

Barragán, R., coord.
2003 *Guía para la formulación y ejecución de proyectos de investigación* (3ra. ed.). La Paz: PIEB.

Berenson, M. y D. Levine
1992 *Estadística básica en administración. Conceptos y aplicaciones*. México: Prentice Hall.

Cea D´Ancona, M. A.
1999 *Metodología cuantitativa. Estrategias y técnicas de investigación social*. Madrid: Síntesis.

Carmines, E. y R. Zeller
1979 *Reliability and validity assessment*. Beverly Hills, Sage.

Chambers, J. M. *et al.*
1983 *Graphical methods for data analysis*. Boston: Duxbury Press.

Cochran, W. G.
1981 *Técnicas de muestreo*. México: CECSA.

Fowler, F. J.
1988 *Survey research methods.* Beverly Hills: Sage.

Freund, J. y G. Simon
1994 *Estadística elemental*. México: Prentice Hall.

Goode, W. J. y P. K. Hatt
1975 *Métodos y técnicas de las ciencias sociales*. Barcelona: Hispano Europea.

Hernández, R., C. Fernández y P. Baptista
1999 *Metodología de la investigación*. México: McGraw Hill.

INE
2001 *Programa de mejoramiento de las encuestas y Medición de las Condiciones de Vida* (MECOVI). La Paz: INE.

Lazarsfeld, P.
1973 "De los conceptos a los índices empíricos", *Metodología de las ciencias sociales*. Barcelona: Laia.

Lea Plaza, S., X. Vargas y A. Paz
2003 *Tarija en los imaginarios urbanos. Investigaciones regionales*. La Paz: UAJMS/CED/CER-DET/PIEB.

Merton, R. K.
1983 *Teoría y estructuras sociales*. México: FCE.

Mills, R.
1985 *Estadística para economía y administración*. México: McGraw-Hill.

Namakforoosh, M. N.
2000 *Metodología de la Investigación*. México: Limusa.

National Council of Teachers of Mathematics
1975 *Recopilación, organización e interpretación de datos*. México: Trillas.

Neter J., W. Wasserman y G. Whitmore
1992 *Estadística aplicada*. Washington: RTI.

Peña, P., R. Barahona, L. E. Rivero y D. Gaya
2003 *La permanente construcción de lo cruceño. Un estudio sobre identidad en Santa Cruz de la Sierra*. Investigaciones regionales. La Paz: UAGRM/CEDURE/PIEB.

Piña, N. C.
1966 "El análisis y la interpretación de los datos", *Serie Introducción a la metodología de la investigación* No. 4. La Paz: OMS/OPS.

Rand Corporation, The
1955 *A Million Random Digits with 100,000 Normal Deviates*. New York: The Free Press.

Rodríguez, J.
1993 *Métodos de muestreo. Casos prácticos*. Colección *Cuadernos Metodológicos* No. 6. Madrid: CIS.

Salman, T.
2003 "Algunas diferencias entre estudios cuantitativos y cualitativos en estrategia metodológica", *Guía de formulación y ejecución de proyectos de investigación*. La Paz: PIEB.

Sánchez Carrión, J. J.
1995 *Manual de análisis de datos*. Madrid: Alianza.

Sánchez Crespo, J. L.
1971 *Principios elementales del muestreo y estimación de proporciones*. Madrid: INE.

Sierra Bravo, R.
2003 *Técnicas de investigación social. Teoría y ejercicios*. Madrid: Thomson.

Stevens, S. S.
1951 "Mathematics, measurement and psychophysics", *Handbook of Experimental Psychology*. S. S. Stevens, ed. New York: John Wiley& Sons.

Tapia, L., E. Quintana, D. Ance y J. Morales
2002 *Calidad de vida en Potosí. Efectos ambientales en cuatro zonas del municipio*. Investigaciones regionales. La Paz: ISALP/DICYT/ PIEB.

Wallace, W. L.
1976 *La lógica de la ciencia en la sociología*. Madrid: Alianza.